Integration of Genome Scale Data for Identifying New Biomarkers in Colon Cancer

Integrated Analysis of Transcriptomics and Epigenomics Data from High Throughput Technologies in Order to Identifying New Biomarkers Genes for Personalised Targeted Therapies for Patients Suffering from Colon Cancer

Aamir Ul HASSAN

Submitted for the Degree of

Doctor of Philosophy

Faculty of Engineering and Informatics

University of Bradford

2017

ABSTRACT

Keywords: Colon Cancer, Microarray Gene expression profiling, Gene ontology enrichment analysis, MicroRNA, System Biology, Bioinformatics, Gene signature, Cross-Validation, Diagnostic and Prognostic.

Colorectal cancer is the third most common cancer and the leading cause of cancer deaths in Western industrialised countries. Despite recent advances in the screening, diagnosis, and treatment of colorectal cancer, an estimated 608,000 people die every year due to colon cancer. Our current knowledge of colorectal carcinogenesis indicates a multifactorial and multi-step process that involves various genetic alterations and several biological pathways. The identification of molecular markers with early diagnostic and precise clinical outcome in colon cancer is a challenging task because of tumour heterogeneity.

This Ph.D.-thesis presents the molecular and cellular mechanisms leading to colorectal cancer. A systematical review of the literature is conducted on Microarray Gene expression profiling, gene ontology enrichment analysis, microRNA and system Biology and various bioinformatics tools.

We aimed this study to stratify a colon tumour into molecular distinct subtypes, identification of novel diagnostic targets and prediction of reliable prognostic signatures for clinical practice using microarray expression datasets. We

performed an integrated analysis of gene expression data based on genetic, epigenetic and extensive clinical information using unsupervised learning, correlation and functional network analysis. As results, we identified 267-gene and 124-gene signatures that can distinguish normal, primary and metastatic tissues, and also involved in important regulatory functions such as immune-response, lipid metabolism and peroxisome proliferator-activated receptors (PPARs) signalling pathways.

For the first time, we also identify miRNAs that can differentiate between primary colon from metastatic and a prognostic signature of grade and stage levels, which can be a major contributor to complex transcriptional phenotypes in a colon tumour.

.

Dedicated to

Who didn't remain with me in life but lives in my heart...

ACKNOWLEDGEMENTS

I would like to extend my appreciations to the number of people who been so generously assisting and contributed to the work required for the preparation of this thesis.

My special thanks goes to my enthusiastic supervisor, Dr Yonghong Peng. My PhD has been a wonderful experience and I am thankful to Dr Yonghong Peng wholeheartedly, not only for his incredible support through the tough time of my studies but also morally lifting me when I was down academically and also other matters of practical life. I thank Dr Yonghong Peng for giving me so many wonderful opportunities.

Similar, profound gratitude goes to Professor Demetres Kouvatsos, who has been a truly dedicated mentor. I am particularly indebted to Professor Demetres Kouvatsos for his constant faith in my research work, and for his support.

I am also hugely appreciative to Masood Zaka, especially for sharing his expertise so willingly, and for being so dedicated as a group member.

Finally, but by no means least, thanks go to mum, dad and my wife for almost unbelievable support. They are the most important people in my world and I dedicate this thesis to them.

TABLE OF CONTENTS

LIST OF FIGURES

List of Abbreviations

The following table describes the significance of various abbreviations and acronyms used throughout the thesis.

Abbreviation	: Meaning
8-oxoG	8-oxo-7,8-dihydroguanine
ACF	Aberrant crypt focus
ALL	Acute lymphoblastic leukaemia
AIRPOLIFE	Air Pollution in a Life-time Health Perspective
AML	Acute yeloid leukaemia
AP	Apurinic/apyrimidinic
APC	Adenomatous polyposis coli
APE1	AP endonuclease 1
ASE-1	Antisense ERCC1
BER	Base excision repair
CC	Colon/Colorectal cancer
cDNA	Complementary DNA
CI	Confidence interval
COX-2	Cyclooxygenase-2
CRC	Colorectal cancer
DCC	Deleted in colorectal carcinogenesis
DCH	Diet, Cancer and Health
DNA	Deoxyribonucleic acid
dRp	Deoxyribose phosphate
DSB	Double-strand break
EPIC	European Prospective Investigation into Cancer and Nutrition
ERCC1	Excision repair cross complementary group 1
FAP	Familial adenomatous polyposis

FEN1	Flap endonuclease 1
GEP	Gene expression profiles
GG-NER	Global genome nucleotide excision repair
GPX	Glutathione peroxidase
HCA	Heterocyclic aromatic amines
HRT	Hormone replacement therapy
IARC	International Agency for Research on Cancer
IRR	Incidence rate ratio
KAM	Kolorektal cancer, Arv og Miljø
LOH	Loss of heterogeneity
MAPK-ERK	Mitogen activated protein kinaseextracellular signal-regulated kinase
MeSH	Medical Subject Heading Terms
MLH	MutL homologue
MMR	Mismatch repair
mRNA	Messenger RNA
MSH	MutS homologue
MSI	Microsatellite instability
NCBI	National Center for Biotechnology Information
NER	Nucleotide excision repair
NF-\B	Nuclear factor-kappa B
NOC	N-nitroso compounds
NORCCAP	Norwegian Colorectal Cancer Prevention study
NSAID	Non-steroid anti-inflammatory drugs
OGG1	8-oxoguanine glycosylase 1
OR	Odds ratio
P53	Protein 53
PAH	Polycyclic aromatic hydrocarbons
PARP-1	Poly (ADP-ribose) polymerase 1
PNK	Polynucleotide kinase

Pol	Polymerase
PPAR	Peroxisome proliferator-activated
PPAR	Peroxisome proliferator-activated gamma
RAI	RelA-associated inhibitor
RF-C	Replication factor
CRHOA	Ras homolog gene family, member A
RNA	Ribonucleic acid
ROS	Reactive oxygen species
RPA	Replication protein A
SNP	Single nucleotide polymorphism
SSB	Single-strand break
TC-NER	Transcription-coupled nucleotide excision repair
TFIIH	Transcription factor II H
TGF-U	Transforming growth factor U
UNG	Uracil DNA glycosylase
UTR	Un-translated region
XPA	Xeroderma pigmentosum complementation group A
XPC	Xeroderma pigmentosum complementation group C
XPD	Xeroderma pigmentosum complementation group D
XRCC1	X-ray cross-complementing 1

CHAPTER 1 INTRODUCTION

1.1 Colon Cancer

Colon cancer (CC) is the most common malignancies in the world and accounts for about 10% of all cancer deaths in both Europe and the USA. (Perez Villamil et al., 2012b). Regardless of advances in screening methods, diagnosis, and therapies, colorectal cancer is the third most common cancer and the fourth-leading cause of cancer death worldwide (Greenlee et al., 2000). Till date, histopathological staging is the only prognostic classification method is used in clinical practices for the selection of patients for chemotherapy treatment. However, it has often occurred that cancer staging on the basis of pathological prognosis fails to predict recurrence accurately in many patients undergoing curative surgery for localized CC. In fact, 10%–20% of patients with stage II CC, and 30%–40% of those with stage III CC, develop recurrence (Marisa et al., 2013a). Extensive investigative studies on colon cancer for the discovery of molecular markers for characterization and prognosis revealed that microsatellite instability (MSI) is caused by defective function of the DNA mismatch repair (MMR) system and the only marker that was reproducibly found to be a significant prognostic factor in early CRC in both a meta-analysis and a prospective trial (Hutchins et al., 2011, Popat et al., 2005). Microarray gene expression profiling is a powerful tool for the identification of diagnostic and prognostic gene signatures. Supervised analysis of gene expression has been used to discover gene signatures to identify patients at risk of recurrence of colon cancer (Zlobec et al., 2010, Zhang, 2008, Yamasaki et al., 2007, Wang et al., 2004, Tran et al., 2011). Therefore, a number of studies have implemented microarray technology to investigate gene expression profiles (GEPs) in CC in

recent years, but unfortunately none of the well-established gene signatures have been discovered that could be beneficial in clinical practice, especially for predicting clinical outcome (O'Connell et al., 2010, Eschrich et al., 2005). For the precise drug targets, molecular homogeneity may be essential in order to identify specific biological pathways affected. Gene expression profiling based studies on CC have been only poorly reproducible largely because CC disease is composed of distinct molecular entities that may develop through multiple pathways on the basis of different molecular features (Kang, 2011, Jass, 2007, Marisa et al., 2013a).

1.2 Role of Microarray in Cancer Research

Cancer is a clinically heterogeneous disease. During the past century, the clinical behaviour of various cancers was determined using histopathology analysis, a process that often lacks exact severity of cancers. Therefore, the microscopic approach can only predict general categories of cancer and cannot reach high specificity and sensitivity prediction in clinical practice (Liotta and Petricoin, 2000). So there is consistently need of novel methods and tools which can answer the heterogeneity of these tumours and complement histopathological evaluation to increase the specificity and sensitivity in cancer diagnosis and prognosis.

The Central dogma of molecular biology elucidates that gene expression is a two-step process in which genetic information is first transcribed into messenger RNA (mRNA) through transcription process followed by translation process into a fully functional form known as proteins, which are also a major structural component of the cell. There is strong understanding that all major cellular process is controlled by certain collective expressed genes, therefore, it is a great value to

analyse genome-wide mRNA level (Brown and Botstein, 1999). Advances in technology along with completion of Human Genome Sequence has offered new technology, DNA microarray technology. DNA microarrays have the ability to simultaneously analyse thousands of mRNA (gene expression level) genes in a given sample. The methodology used in this type of measurement generally known as gene expression profiling. The introduction of DNA microarray technology provided much-improved understanding of various cancers and allowed researchers to analyse thousands of cancer genes along with their role in disease progression and their expression patterns linked to clinical phenotypes (Lonning et al., 2005). The DNA microarray technology also offers great benefits in terms of identification of gene signatures who have the ability to differentiate cancer from normal and metastatic tissues, capable of predicting outcomes and recurrence, and response to treatment. The technology also offers a great deal of potential for discovery of novel drug targets and improve our understanding of the disease causes and progression.

1.2.1 DNA Microarray

The idea behind DNA microarrays is a specific hybridization of complementary nucleic acid sequences between the two DNA fragments, one is DNA microscopic spots attached to the solid surface and other is fluorescent labelled RNAs from target tissue (Southern et al., 1999). Generally, a DNA microarray consists of thousands number of DNA spots, known as probes or reporters or oligo, on a glass. The signal intensity of hybridization of each probe-target is quantified by detection of fluorophore should correlate to the abundance of mRNA in a target. There are generally two types of DNA microarrays based on the DNA fragment used in the constitution of arrays, oligonucleotide arrays and

complementary DNA (cDNA) arrays. In oligonucleotide arrays, probes are synthesizing directly on the surface of silicon water whereas cDNA array is prepared using polymerase chain reaction amplification of interesting genes from cDNA library.

1.2.2 DNA Microarray Analysis

A number of advance tools have been proposed recently for the analysis and interpretation of DNA microarrays data. These tools are commonly divided into two major categories, supervised learning (classification) and un-supervised (clustering) learning methods. The goal of supervised learning methods is to identify gene expression patters e.g. gene expression signature, which can be used for the classification of unknown samples according to their biological or functional characteristics (Golub et al., 1999). So, supervised learning methods provides an opportunity to integrate knowledge of class label into analysis for the classification of samples. Hence, in a given microarray data set consisting of gene expression matrix and class label, a subset of most diagnostically discriminatory genes can be selected by building a predictive model, also known as classifiers e.g. SVM (support vector machine) and *K*-NN (*K*- nearest neighbours). The underlying concept of these classifiers is to take input expression matrix of the pre-selected set of genes of unknown samples as an input and predicts the class label of each sample. Supervised learning is largely used for two class problems e.g. cancer versus normal, or multi-class problems such as identifications of sub-types of same cancer.

In contrast, unsupervised learning methods involve information of samples, genes or both into different clusters consist of similarities in gene expression values (Eisen et al., 1998). The fundamental goal of clustering is to divide

objected into groups with similar characteristics. Several clustering has been applied to microarray data analysis such as hierarchical clustering, self-organising maps, and *K*-means clustering (Quackenbush, 2001). The advantage of unsupervised learning is that this is method is unbiased and allows significant discoveries in complex data sets without any background information about the structure.

1.3 DNA Microarray and its Role in Cancer Research

DNA microarray technology has been extensively used in cancer research for the past numbers of years now, specifically in search of gene expression signatures for the prediction of diagnostics and prognostics categories of cancer patients (Sørlie et al., 2001, Golub et al., 1999).

1.3.1 Diagnostics Classification

One of the most prevalent and challenging problem in cancer research is the histopathological identification and classification of various cancers. Similar morphological cancers may belong to distinct clinical subtypes despite the same origin. The ability to identify unknown cancer samples into particular subclasses may provide an edge for more efficient cancer diagnosis.

The earliest study by Golub *et al.* (Golub et al., 1999) exploited the gene expression profiling technique for cancer diagnosis of 6, 817 genes in 72 human acute leukaemia tumour samples. They used unsupervised learning method and suggested two major clusters of leukaemia subtypes, acute myeloid leukaemia (AML) and lymphoblastic leukaemia (ALL). Followed by the clustering into subtypes, a weight gene classifier was build using supervised learning method for the classification of unknown samples into the correct class of AML and ALL.

The accuracy of the classifier was assessed using cross-validation on training set as well tested on the independent set of samples.

A number of other studies have also used gene expression profiling techniques for the classification of various cancers (Bittner et al., 2000, Welsh et al., 2001, Perou et al., 2000, Bhattacharjee et al., 2001). Adding to that, diagnostics classification studies have been performed on number of different cancers such as, prostate cancer (Singh et al., 2002), lung cancer (Bhattacharjee et al., 2001), breast cancer (Perou et al., 2000), bladder cancer (Dyrskjøt et al., 2003), head and neck cancer (Belbin et al., 2002) and ovarian cancer (Ono et al., 2000). For the classification of multiple types of cancers, several techniques of multiple tumour classifiers have been proposed by exploring the microarray gene expression data to discriminate different kind of cancer types based on their tissues of origin (Su et al., 2001, Ramaswamy et al., 2001). These methods includes classification tree (Giordano et al., 2001), linear discriminatory analysis (LDA) (Shen et al., 2006), artificial neural network (ANN) (Bicciato et al., 2003), nearest neighbour classifier (Li et al., 2001), and support vector machine (Liu et al., 2005).

Support vector machine is the most popular classifier applied in microarray data analysis of different cancers. The first application of support vector machine was achieved by Mukherjee *et al.* (Mukherjee et al., 1999) followed by the wide spread use in molecular classification of microarray expression data. Support vector machines classifier were largely designed for two binary classifications but can be customised for multiple tumour types.

1.3.2 Prognostic Classification

One of the most intriguing and contributing applications of DNA microarrays in cancer research is the estimation of clinical outcome based on gene expression profiles. Unlike molecular diagnostics predication, clinical outcome prediction is not just dependent on gene expression profiles but mainly deals with a correlation of gene expression profiles and clinical outcome.

The earliest study conducted by Alizadeh *et al.* (Alizadeh et al., 2000) in diffused B-cell lymphoma (DLBCL) samples, and predicted the correlation between gene expression profiles and clinical outcome. This study uses unsupervised learning method on B-cell malignancies and identified two molecular forms of DLBCL which indicate different stages of B-cell. This study on DLBCL has led the identification of previously undetected subtypes.

Another major focus of DNA microarray was a prognostic classification of Breast cancer. Studying lymph-node status at the time of diagnosis is one of the best indicators of future relapse and survival outcome of breast cancer patients. Earliest studies have also indicated that the systematic adjuvant chemotherapy reduces the risk of metastasis and also survive breast cancer long (Cole et al., 2001). A pioneering study in clinical outcome prediction was conducted by can't Veer *et al.* (van 't Veer et al., 2002), applied microarray data analysis to primary tumours of 117 patients samples of lymph-node-negative and identified 70-gene prognostic signature predictive of clinical outcome. The prognostics signature was further validated on independent data set of 295 patients samples with best discriminatory power currently observed in the breast cancer clinical prediction.

The successful validation of 70-gene prognostic signature has led the discovery of many other prognostic signatures of breast cancer (Naderi et al., 2007, Wang et al., 2005b, Sotiriou et al., 2006). Similarly outcome prediction based on gene expression profiles was expanded to many other cancer types such as, lung cancer (Beer et al., 2002), prostate cancer (Singh et al., 2002), brain cancer (Pomeroy et al., 2002), and renal cancer (Takahashi et al., 2001). Studies such as above highlights the great potential for gene expression profiling in the identification of prognostic signatures for predicting clinical outcome. However further validation of this gene signature is required.

1.4 Microarray Data Integration

Recent studies highlighted the discovery of DNA microarray provided powerful tools which accomplished meaningful insights in cancer research. Similarly other studies have also utilised this technology to identify vital gene signature that can discriminate between the diagnostic categories along with the prognosis of cancer patients (Sotiriou et al., 2006, Bittner et al., 2000, Sørlie et al., 2001, Eisen et al., 1998, Golub et al., 1999, van 't Veer et al., 2002). However, microarray characteristic of high noise, cost and small patient samples size of individual study makes it difficult to identify reliable gene signature for diagnostics.

Another interesting but anticipated evidence is that only a small overlap between the gene signatures have been observed from the studies of a different investigator of same cancer type, may be due to protocol differences. A recent study by Ein-Dor *et al.* (Ein-Dor et al., 2005) question this disagreement on signature uniqueness is not just because of different platform, sample scarcity and experiment protocol, but the signature prediction is strongly influenced by a

subset of the patients used for the signature identification. Hence, the signature difference might be largely contributed by sample size in individual studies and suggested large sample cohort will produce more robust signatures.

Considering the cost of performing DNA microarray analysis and scarcity of certain tumour samples such in our case (colon cancer), it is very difficult to performed large scale analysis. In addition to that, it is very hard to repeat expression reading from the valuable specimens. Therefore, accumulation of gene expression data achieved from various investigators, generated in different laboratories may address the question surrounding small sample size. Another advantage of this microarray data integration would be to identify gene features which might be covered by small samples size and experimental protocol. Successful integration might lead to the discovery of robust and accurate gene signature important for diagnostics classification and improve statistical correlation with clinical outcome.

1.4.1 Types of Microarray Data Integration

The rapid increase in microarray data has offered researcher to proposed novel integration methods which can effectively accumulate data generated by various investigators and from different laboratories. In principle, integration of multiple microarray platform data should produce more reliable results since the analyses are performed on large samples size and individual study biased is also reduced. Several methods have been proposed for inter-study microarray data integration at the different level of studies. But these methods can be divided into two major categories considering their level of integration: meta-analysis, which combined the results after the class comparison (e.g. t-test statistics) from individual data sets which also avoid direct comparison of expression matrixes, and direct

integration of individual expression data matrixes after performing quality control tests such as transformation and normalization.

1.4.2 Problems in Data Integration

In ideal condition, any gene expression experimental data obtained from any research laboratory, using any protocol, at any time, using microarray technology platform, should be comparable. However, this is unachievable in reality and often poses massive challenges due to lack of uniformity in standards to microarray data integration. Numerous problems have been identified when attempting to integrate microarray data generated by individual studies groups using multiple array platforms.

Till now, several studies have shown that gene expression level measurements from different microarray platforms, such as spotted cDNA and oligonucleotide arrays, might have poor correlation and direct comparison might be meaningful (Tan et al., 2003, Mah et al., 2004, Kuo et al., 2002). The identifying divergence among the different arrays may be due to the differences in probe set used, platforms technologies, labelling and hybridizing protocols, as well as differences in data extraction techniques such as background correction, normalization, and calculation of expression values. Such as, the data obtained from cDNA microarray is usually defined by the ratio between the diseases (experimental) and the control expression values which cannot be directly compared with the oligonucleotide microarray data which is defined as expression values of disease samples.

Along with the observed difference between the various microarrays platforms, comparison between the microarrays data obtained from multiple-laboratories

have demonstrated major differences in data from individual studies extracted using the same microarray platforms than that obtained in the same lab using different microarray platforms (Nimgaonkar et al., 2003, Wang et al., 2005a). This particular issue also suggests that data obtained from different laboratories could not be compared even though the data was extracted using same microarray platform.

Another issue linked to microarrays is a broad lack of reproducibility among the generations. Commercially produced microarrays such as Affymetrix have several generations of microarrays in order to compete with the advances in gene sequencing. So any new discoveries related to novel genes and their representative composition is frequently shared and updated with new developments microarrays using the field of biotechnology. Consequently, any probe set data consisting of newly discovered genes are incorporated into new generations of commercially available microarrays and the existing probe sets are modified for better detection targeted gene sequences. A recently conducted study has highlighted the issue of reproducibility among the two different generations by analysing the reproducibility factor. They identify the Affymetrix are high reproducible with in one generation but reproducibility across the two different generation depend up on the degree of similarity among the probe sets and the expression levels (Nimgaonkar et al., 2003). This issues suggest even using the same microarray platform but having different generations would make direct integration difficult due miss-matching in the probe sets and duplicated spots.

In addition to above-discussed issues, there is multiple factors which could make the direction comparison very difficult. Such as, variation among the data sets, the difference resulting from the technical variability, the difference in sample size,

preparation methods and experiments controls, array quality, RNA detecting methods and RNA quality further deepens the challenges to integrating data from individual's studies.

1.5 Thesis Overview

The aim of this study is to performed microarray integrated analysis on high-throughput colon cancer data for the prediction of novel diagnostic and prognostic signature identification.

The main objective of this thesis was to propose novel diagnostic and prognostic signature gene which can address the issue of heterogeneity and accurately predict the clinical outcomes of patients when tested on the wider population. As we highlighted earlier that many attempts have been made for the prediction of colon cancer signatures that can precisely separate colon cancer from others and accurately predict the outcome in clinical practices. Chapter 3 presents the first independently study where we implemented tissue-based integration analysis and identified novel diagnostic and prognostic signatures by establishing workflow

Independent of microarray data limitation of sample size, platform and other experimental protocols. We searched the PubMed database for the latest development in colon cancers applying the microarray tools and selected the high-quality studies. We performed the tissue-based comparison for the selection of differentially expressed genes and the searched for the overlapping genes in all the comparison. We tested the diagnostic signature proposed from the normal versus primary colon comparison on cross-validation loop for their ability to discriminate colon tumours from normal colon among all the cohort data sets. We also tested prognostic signature generated from the comparison of normal versus metastatic

and primary colon versus metastatic colon using univariate cox proportional hazard analysis.

In the second independent study, we performed integration analysis of two different data levels for the identification of factors which contribute the heterogeneity of colon cancers and for the identification of functional genes involved in the metastases of colon cancer. As we know, in order to understand the genetic basis of complex traits such as colon cancer has been a challenge for past few years now. But with the increase in expression data and advances in technology from multiple levels of biological systems has been a game changer. Various analytical approaches have been developed to identify the genetic variation that underlies complex traits. Such as variation in gene expression (using microarrays and RNA sequencing (RNA-seq)), epigenetic variation (by microRNAs, methylation arrays, methylation sequencing or chromatin immunoprecipitation followed by sequencing (ChIP–seq)) and protein variation. Therefore, in this study, we identified microRNA signature that can differentiate primary and metastatic tumours, its stages and tumour grades in patients with colorectal carcinoma. In an integrative model, we performed differential expression analysis between histopathological groups and identified significantly dysregulated miRNAs followed by the target prediction analysis to establish their tumour transcriptional phonotypes. Furthermore, we focused on the identification and evaluation of miRNAs potentially involved in prognosis and functional processes during metastatic progression.

CHAPTER 2 LITERATURE REVIEW

2.1 Background of Colon Cancer

Numerous of studies have confirmed the use of microarray data analysis role of expression profiling in the identification of novel diagnostic markers, drug targets for personalised medicine and tumour classifications into subtypes. In 2002, the earliest effort of an integrated analysis of data consisting of gene expression profiling and drug sensitivity have been used in the assessment of clinical tumours for the prediction of biomarkers that can predict therapeutic efficacy (Orr and Scherf, 2002). This technique of gene-drug correlation had been employed for selecting therapeutic options for tumours on the basis of their molecular characteristics. The main focus of this technique was sensitivity rather than actually focusing on molecular significances of therapy. Therefore, the earliest use of gene expression data in correlation with drug sensitivity database was a proposal of antimetabolite 5-FU gene-drug correlation, which was used for the treatment of breast and colon cancers by inhibiting both RNA processing and thymidylate synthesis.

Another way of studying the development of any type of a tumour is by closely analysing the sequence of events. Studying this general model for colon cancers and all those sequences that lead formation of adenoma to carcinoma usually occur through genetic and epigenetic changes (Fearon, 2011). Similarly, various molecular phenotypes have been used to classify the colon cancers in the past (Sanchez et al., 2009). Such as phenotype based on microsatellite instability (MSI) (Iacopetta et al., 2010), phenotypes based on epigenetic changes were studying of methylation state of CpG islands (van Engeland et al., 2011), phenotypes based on the genetic aberrations were presence of genes such as KRAS or BRAF, and

phenotypes based on functional pathways were presence of Wnt/ß-catenin, TGF-ß, MAPK, and PI3K signaling (Fearon, 2011).

In the past, molecular studies mainly focused on individual gene targets rather than addressing the molecular heterogeneity of the disease. The introduction of DNA microarray technology offered an investigation of thousands of genes at the mRNA expression levels. Therefore, provided an alternate to single gene discovery to thousands of genes simultaneously in a single assay, thus offered solutions to address the problem of heterogeneity of complex diseases (Mohr et al., 2002, Bertucci et al., 2001). Number of earlier publications based on gene expression profiling of colon cancers have performed comparison of normal versus tumour tissue samples or comparison among the histological stages (Alon et al., 1999, Backert et al., 1999, Hegde et al., 2001, Kitahara et al., 2001, Notterman et al., 2001, Agrawal et al., 2002, Birkenkamp-Demtroder et al., 2002a, Lin et al., 2002, Zou et al., 2002, Frederiksen et al., 2003b, Tureci et al., 2003, Williams et al., 2003) but none of them focused on the disease prognosis and MSI phenotype. Bertucci and colleagues (Bertucci et al., 0000) have used application of DNA microarray data analysis on 8,000 genes of 50 colon cancer tissues and identified several dysregulated genes among the normal and cancer samples along with the prediction of clinical outcomes and MSI phenotype.

Molecular level similarities are essential in the identification of specific pathways affected in diseases, in the identification drug targets or achievement and designing of survival outcome classifiers.

In 1932 the British pathologist Cuthbert Dukes devised a classification system for colon cancer. Several different forms of the dukes' classification were developed. Such as:

Dukes'A: The cancer is in the inner linning of the bowel, it is slightly growing into the muscle layer.

Dukas's: cancer has grown through the muscle layer of the bowel.

Dukes'C: cancer has spread to at least 1 lymph node close to the bowel.

Dukes'D: cancer has spread to another part of the body, such as liver, lungs or bones

Previously, several attempts have been made to subdivide colon tumours into further sub-classes or to correlate gene expression with Dukes stages hasn't been fruitful. Some of them successfully classify normal colon, Duke B and C but not Duke A and D (Frederiksen et al., 2003a), one of them has been able to classify Duke A in normal colon but not B, C and D (Birkenkamp-Demtroder et al., 2002b). Other authors were unable to find the difference between stages of colon A, B and C (Kwong et al., 2005). Perez Villamil et al. proposed a novel method aiming to obtain more homogeneous groups of colon tumours for the discovery of molecular uniform subgroups that are more likely to discriminate patient with different clinical outcomes along with the better understanding of biological pathways forming different tumour subtypes (Perez Villamil et al., 2012a).

In another study conducted by a group of Cancer Genome Atlas performed genome-scale analysis of 276 patients suffering from colon cancers (The Cancer Genome Atlas, 2012). In this multidimensional approach, the investigators divided

colon data into those with microsatellite instability (MSI) and those that are microsatellite-stable (MSS). In results, they observed a number of important genes and critical pathways required for the initiation and progression of colon cancers. Some of the significant findings from this study were the discovery of P53, PI3K, RAS-MAPK, TGF-β, WNT, and DNA mismatch repair pathways. In spite of such progress, there are still some unknown genetic and genomic changes which play a significant role in colon tumorigenesis.

Gene expression profiling-based on colon cancers have been poor in terms of reproducibility largely due to involvement complex pathways and heterogeneity with in disease (Kang, 2011, Jass, 2007). As a result, several diagnostic and prognostic signatures have been proposed addressing the distinct gene features and pathways. Recently, gene expression profiling-based studies have integrated genetic and epigenetic analysis has identified three distinct molecular subtypes of colon cancers (2012, Jass, 2007, Salazar et al., 2011). Hence colon cancer should not be considered as one disease but a collection of sub-entities. However, there is urgent need of redefining molecular classification currently based on biomarkers such as MSI, CpG island methylator phenotype CIMP, chromosomal instability CIN, and BRAF and KRAS mutations (Shen et al., 2007, Kang, 2011).

2.2 Review of Microarray Data Integration

The fast pace increase in microarray data has forced the researcher to create and build effective methods to integrate data produced from disparate laboratories using different microarray technology platforms. There is a strong agreement in the integrating multiple data sets to produce more reliable and strictly valid results. This is largely due to the reasons that yield analyses are performed using a

larger number of samples and the effects of individual study-specific biases are reduced. Consequently, a number of methods have been introduced to integrated inter-study microarray data at different levels in microarrays (Warnat et al., 2005, Ghosh et al., 2003, Ramaswamy et al., 2003). These integration methods broadly fall into two categories: a meta-analysis, which combines the results (t-test statistics) from individual studies just to avoid the direct comparisons of expression values, and direct integration of expression values after some transformation methods such as normalization etc.

Several studies have adopted this method of combining results of individual studies to increase the power of identifying significantly expressed genes across studies instead of integrating microarray gene expression values (Zhou et al., 2005, Stevens and Doerge, 2005, Ghosh et al., 2003, Ramaswamy et al., 2003, Nimgaonkar et al., 2003). One of the earliest efforts by Rhodes et al. (Rhodes et al., 2002) in which they exploited the meta-analysis method and proposed a statistical model for integrating four independent microarray data sets from two different microarray platforms, spotted cDNA arrays and Affymetrix arrays, respectively. Each of the identified gene in each study was treated as an independent hypothesis and significance value (p-value/q-value) was allocated to each gene in each study based on random permutations. Following to above step, the similarity of significance across studies was evaluated with meta-analysis methods combined with multiple inference statistical test for each possible combination of studies. Another study based on significantly dysregulated in prostate cancer; Choi et al. (Choi et al., 2003) proposed a new meta-analysis method, which combines the results from an individual study in the form of effect size and has the ability to model the inter-study variation. The effect size was

calculated mean difference between cancer and normal samples in a microarray data set. Afterword, the author combined the individual study effect sizes from multiple microarray data sets to estimate the overall mean following by calculation of statistical significance using permutation test extended to multiple data sets. They successfully established that combining data using this method supported the discovery of small but consistent expression changes and may increase the sensitivity and reliability of analysis. Hu et al. (Hu et al., 2005) proposed an extended version of effect size model for the meta-analysis of microarray data.

Multiple studies have reported success using meta-data analysis method for integration individual microarray studies but small sample size, coupling with divergences due to the difference in studies protocols have affected the final results of meta-analysis method. This is the main weakness of using this method reported till now. In addition to this, recently studies also highlighted that there is moderate overlap between the gene signature detected using individual studies using different platforms. So there are strong chances of losing important genes using this meta-analysis method (Mah et al., 2004).

CHAPTER 3 INTEGRATED ANALYSIS OF GENE EXPRESSION DATA FOR COLON CANCER BIOMARKER DISCOVERY

3.1 Introduction

Colon Cancer (CC) is a common malignancy affecting both women and men. Despite recent advances in the screening, diagnosis, and treatment of colorectal cancer, an estimated 608, 000 people die every year from this form of cancer. Pathological staging is the only prognostic classification used in clinical practice to select patients for adjuvant chemotherapy. However, pathological staging fails to predict recurrence accurately in many patients undergoing curative surgery for localized CRC. In fact, 10%-20% of patients with stage II CRC, and 30% - 40% of those with stage III CRC develop recurrence (Zhang et al., 2001).

Among the molecular markers that have been extensively investigated for colon cancer (CC) characterization and prognosis. DNA mismatch repair (MMR) system, is the only marker that was reproducibly found to be a significant prognostic factor in early CRC in both a meta-analysis and a prospective trial (Nannini et al., 2009, Zhang et al., 2001). Precise classification of the tumor is critically important for cancer diagnosis and treatment. During the past decade, efforts have been made to use gene expression profiles to improve the precision of classification, with limited success (Cardoso et al., 2007). Many studies have exploited the use of microarray technology to investigate gene expression profiles (GEPs) for the diagnosis of colon cancer in recent years, but no signature has been to be useful for clinical practice, especially for predicting prognosis (Sagynaliev et al., 2005). It is shown that the reproducibility of GEP studies on colon cancer has not been sufficient for clinic practice, possibly because colon cancer cells are composed of distinct molecular entities that may

develop through multiple pathways (Chan et al., 2008, Shih et al., 2005). Therefore, there may be several prognostic signatures for CRC, each corresponding to a different entity.

Indeed, GEP studies, based on integrated analysis of genetic/epigenetic data including high-throughput methylene data (Nannini et al., 2009), have identified at least three distinct molecular subtypes of colon cancer. Therefore, colon cancer should no longer be considered as a homogeneous entity. However, the molecular classification of CC currently used, which is based on a few common DNA markers (MSI, CpG island methylator phenotype [CIMP], chromosomal instability [CIN], and BRAF and KRAS mutations) (Kang, 2011, Jass, 2007), needs to be refined, and a standard and reproducible molecular classification is still not available.

In order for identifying more robust diagnostic gene signature of colon cancer, this Research thesis presents an investigated analysis of multiple latest competitive studies of various stages of colon cancer. We applied tissue-based differential expression followed by supervised machine learning approach for the discovery of diagnostic/prognostic gene signatures for the earlier and outcome identification of patients with colon cancers. We identified a 124-gene signature that can discriminate between the patients with good and poor outcomes, also provide evidence of functionally involved in immune response, lipid metabolism and PPAR signalling pathways.

Table 1. A summary cohort of studies involving Colon cancer.

Author	Studies	Cohort	Sample Size	Platform	Dataset
Musella et al. (2013)	Time course analysis of colon cancer samples	Normal vs Tumor	N=88, T=84	GPL6947 Illumina HumanHT-12 V3.0 expression bead chip	GSE37182
Shaffer et al. (2009)	Expression data from colorectal cancer patients	Normal vs Tumor vs Mets	N=54, T=186, M=67	Affymetrix Human Genome U133A Array	GSE41258
Agesen et al. (2013)	Specific extracellular matrix remodelling signature of colon hepatic metastases	Normal vs Tumor vs Mets	N=18, T=20, M=19	Affymetrix Human Genome U133A Array	GSE49355

This table is just showing the numbers of studies included in our analysis cohort. It can clearly be identified from the table that data sets were derived from different platforms of microarray studies. The first study comparing the normal versus tumour samples, and the other two studies comparison include expression from three different tissues including normal, primary and Mets. The dataset column includes the accession numbers of GEO expression data sets and general from starts from GSE followed by identification number.

3.2 Method

We performed two staged integrated analyses on the expression data sets on three lately developed studies based on gene expression analysis of colon cancer for the discovery of potential gene signature. In order to extract the biological information, we further performed gene ontology enrichment analysis in order to

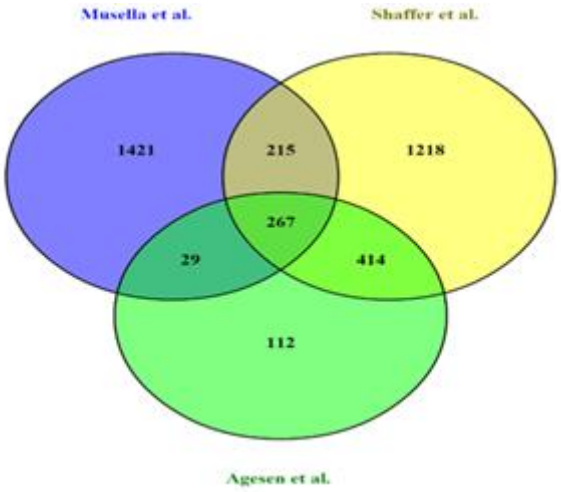
identify the functional pathways involved in localised colon cancer as well as spread to other tissues. We searched studies involving applications of gene expression profiling on patients involving samples primary and metastatic tumour tissues.

3.2.1 Data Collection

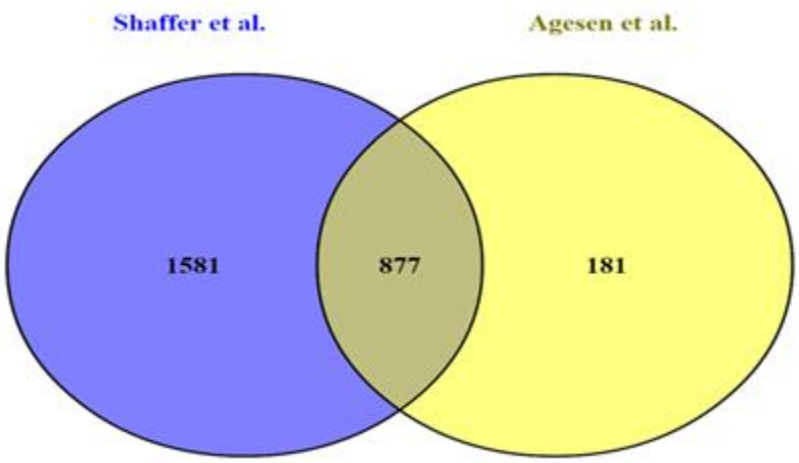
The three microarray expression data sets were obtained using GEO query Bioconductor R package (Davis, 2013) from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo) (Table I). The expression data sets involve samples from normal, primary tumour and metastatic tissue samples. In order to identify tissue specific mRNA signatures, we performed comparisons of among three tissues to identify specific dysregulated mRNAs.

3.2.2 Statistical Analysis

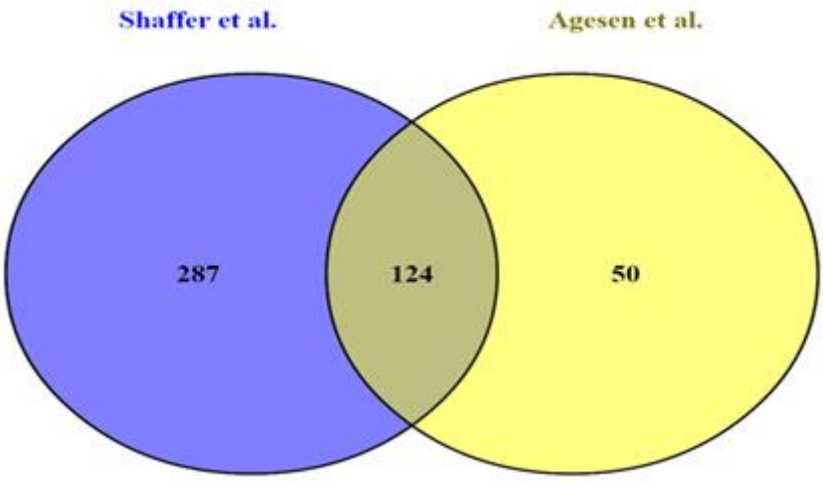
Each of the extracted raw expression data sets was log-transformed and normalised by quantile method individually. Using R/Bioconductor, linear models for microarray data analysis were employed by forming contrast matrix comparisons for normal vs a primary tumour, normal vs Mets and primary tumour vs Mets. Significance value (P-Value < 0.05) and log scale (log FC > 1 | log FC < -1) was used to rank the genes of interest. Corrections for multiple comparisons were done using false discovery rate (FDR) method. NCBI's original genome annotation was used to obtain gene symbols for probe sets id's.



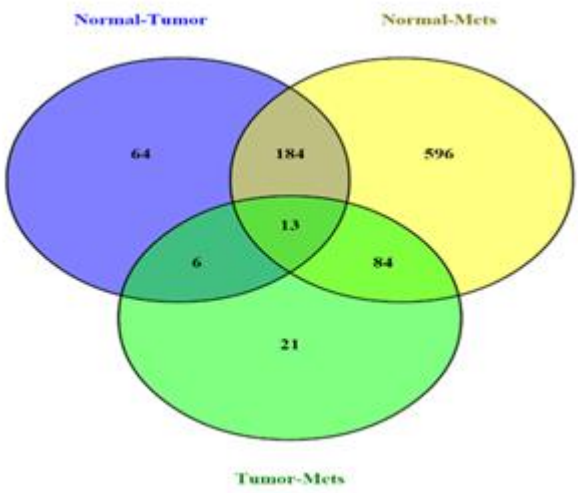
Venn diagram showing the common and unique genes of resulted gene lists after comparison between normal versus tumour samples. We found 267 genes signatures which are common in all three comparisons. Following the identification of dysregulated gene lists from the comparison of normal versus primary tumour tissues classes of each eligible study, we combined the resulted gene lists to form an inter-study signature gene set. In this case, we observed 267 genes were common to three gene lists of normal versus primary tumour comparison of three investigated data sets.



Venn showing normal versus Mets comparison



We focused our attention towards comparison of differentially expressed profiles among a tumour versus Mets tissues comparison and identified 124 genes were common between the resulted gene lists. We observed a number of notable genes were previously linked with metastases in colon cancer.



We further analysed identified signatures by performing meta-analysis among inter-study signature derived from individual comparisons of normal versus a tumour, normal versus Mets and tumour versus Mets tissues by comparing the similarities between them. We observed overlapping of 184 genes between normal versus a tumour and normal versus Mets gene lists. However, 64 genes of normal versus tumour comparison represents a strict tumour-specific (those genes which are not significantly dysregulated in another comparison) pool for which the functional analysis identified their targeted pathways involved.

3.2.3 Functional Analysis

We applied gene ontology enrichment analysis for the interpretation of gene signatures in order to identify potential biological processes, functional network and pathways. For this purpose, we applied Functional Annotation Tool of DAVID Bioinformatics Resources 6.7 (Huang et al., 2009b) using default settings. We used gene symbols as input gene list for each derived signature by selecting the Homo sapiens as their population background. We used P-value < 0.05 as a cut-off value for the selection of DAVID terms and this reason for choosing this criterion so that its conclusions to be drawn about the statistical plausibility and clinical relevance of the study findings.

3.2.4 Classification Performance Evaluation

We applied supervised machine learning approach in order to study the reliability and robustness of inter-study signature. We estimated classification performance on each expression data sets by building a classifier using signature genes, as a feature vector, and their corresponding expression data from (Musella et al. (2013), Shaffer et al. (2009), Agesen et al. (2013)) using the cross-validation loop on support vector machine (SVM) (Mukherjee et al., 1999). We used standard leave one-out cross-validation (LOOCV) to estimate the accuracy of above classifier. Hence, for every sample xn in the training set S, we train the classifier by leaving one sample (N-1) and then classifying the left out sample to predict the label of xn.

3.2.5 Survival Analysis

We performed prognostics analysis for the 124-genes signature derived from the comparison of a tumour versus Mets tissues. For this purpose, we used

independent data set (GSE17538) from the study conducted by Smith et al. (Smith et al., 2010) derived from metastatic colon cancer. We tested 124-genes association with the clinical endpoints such as Overall survival (OS), Disease-specific survival (DSS) and Disease-free survival (DFS) across all the grades (grade 1, 2 and 3) by building Cox proportional hazard (PH) model. We build classifiers using genes from signature genes and their corresponding values from the training set for the calculation of Wald score for each of the genes in the classifier. Log-rank tests P-Value were computed for both univariate and multivariate Cox model for OS, DSS and DFS. Similarly, Kaplan-Meier estimates were calculated for each endpoint.

3.2.6 Validations

For the validation of these studying findings, we applied two approaches: first we validated our proposed signature using in-silico cross validation and second approach we applied to the literature search of signature genes from previously published studies and curated cancer signature databases. Therefore, we searched gene signature database GeneSigDB (http://compbio.dfci.harvard.edu/) for the potential overlaps between the proposed signatures and previously published signatures.

3.3 Results

3.3.1 Gene Expression Analysis and Microarray Data Integration

We performed differential expression analysis on each of the studies by comparing expression profiles of normal, tumour and metastatic tissues samples. We employed t-test statistics among the contrast matrix comparisons for the identification of dysregulated genes.

3.3.2 Expression in Normal and Primary Tumour Colon Tissues

To investigate the difference in human colon cancer, we performed differential expression using normal versus tumour samples of each data set. We identified 2358 dysregulated (2144 up-regulated and 214 down-regulated) genes consisting of 88 normal and 84 primary tumour samples of Musella et al. study. Similarly, we observed 2696 genes (724 up-regulated and 1972 down-regulated) and 1050 genes (366 up-regulated and 684 down-regulated) from Agesen et al. and Shaffer et al. studies, respectively. We excluded all those probes set ids with no gene symbols for further analysis.

Following the identification of dysregulated gene lists from the comparison of normal versus primary tumour tissues classes of each eligible study, we combined the resulted gene lists to form an inter-study signature gene set. In this case, we observed 267 genes were common to three gene lists of normal versus primary tumour comparison of three investigated data sets (Figure 1).

The gene ontology functional analysis of 267-genes shows over-representation of signalling-related molecules in processes and networks (Fig 2). We also identified pathways significantly (P-value < 0.05) involved in various cancers such as bladder cancer and acute myeloid leukaemia. Known signalling and metabolic pathways also featured among the top ten regulatory identified pathways (Table 2).

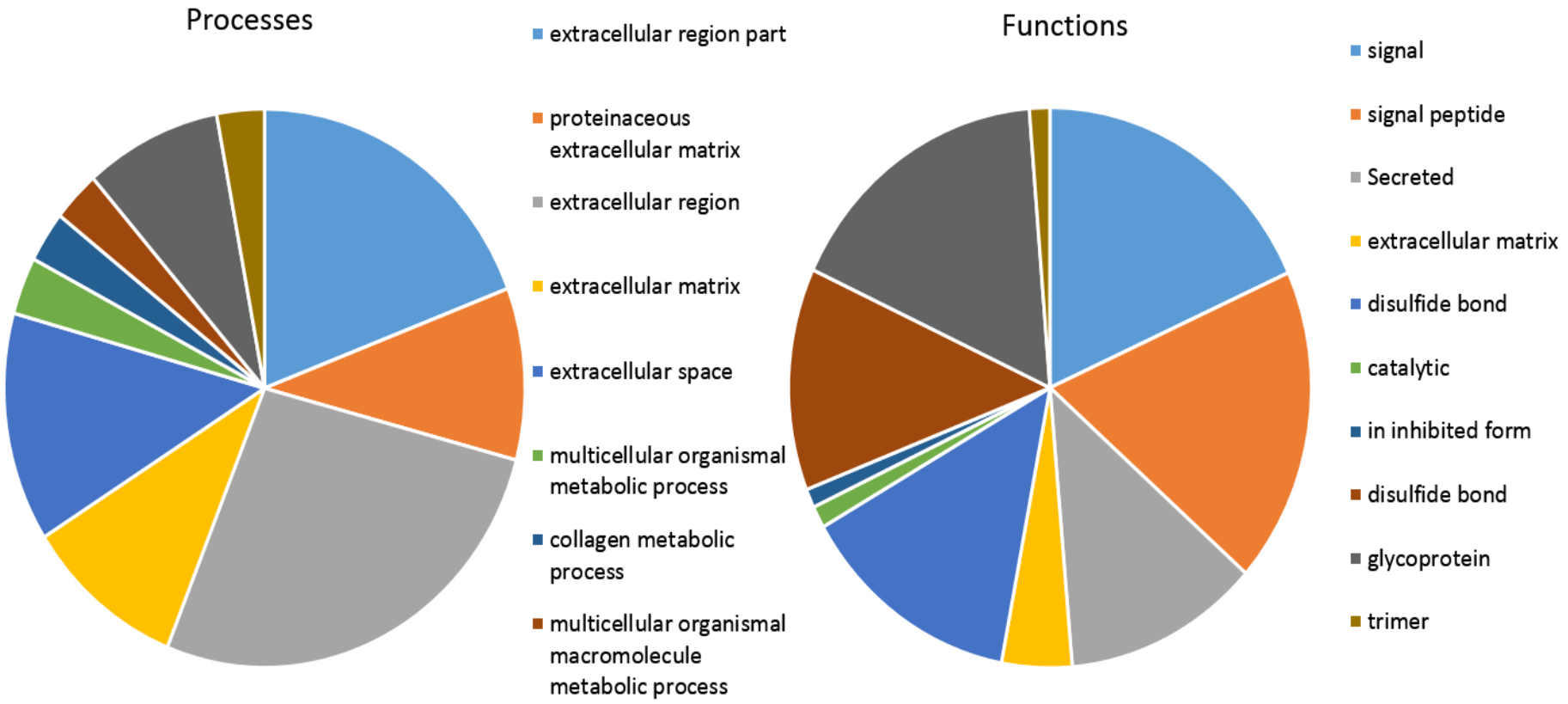


Figure 1. Top-ten processes and molecular functions.

Table 2. Top 10 functional regulatory pathways.

Pathways	P-Value	Genes
Focal adhesion	6.82E-05	CAV1, MET, FLNC, COL5A2, COL5A1, PRKCB, MYL9, CCND1, VEGFA, COL1A2, COL1A1, COL11A1, THBS2, MYLK, SPP1
ECM-receptor interaction	8.12E-03	COL1A2, COL1A1, COL5A2, THBS2, COL11A1, COL5A1, SPP1
Bladder cancer	1.12E-02	CCND1, MMP9, VEGFA, MYC, MMP1
Nitrogen metabolism	1.18E-02	CA7, CA4, CA2, CA1
Complement and coagulation cascades	1.45E-02	C7, F12, CFB, SERPINE1, CFD, PLAU
Cell cycle	1.57E-02	CDK1, CCND1, E2F5, BUB1, MCM4, MYC, CDC25A, CDC25B
Vascular smooth muscle contraction	2.99E-02	KCNMA1, ACTG2, MYH11, KCNMB1, MYLK, PRKCB, MYL9
Acute myeloid leukemia	3.30E-02	CCND1, LEF1, ZBTB16, RUNX1, MYC
Leukocyte transendothelial migration	3.73E-02	CLDN8, MMP9, CLDN1, CXCL12, PRKCB, THY1, MYL9
Wnt signaling pathway	3.90E-02	WNT5A, CCND1, SFRP1, MMP7, CHP2, LEF1, MYC, PRKCB

In table 2 we identified pathways significantly (P-value < 0.05) involved in various cancers such as bladder cancer and acute myeloid leukaemia. Known signalling and metabolic pathways also featured among the top ten regulatory identified.

In the second step of bioinformatics analytics, we investigated reliability and robustness of proposed 264-gene signature using each of the expression data sets generated from different platforms (Table 3). The 264-gene signature consistently achieved high classification accuracy ratios across all the data sets, classifying with 100%, 93.84% and 77.77 %, respectively.

Table 3. Leave one out cross-validation classification of normal versus tumour signature (264-genes).

Author	Expression Set	No. of Samples	Accuracy (%)	Sensitivity (%)	Specificity (%)
Musella et al. (2013)	GSE37182	N=88, T=84	100	100	100
Shaffer et al. (2009)	GSE41258	N=54, T=186, M=67	93.84	92.82	100
Agesen et al. (2013)	GSE49355	N=18, T=20, M=19	77.77	81.7	96.6

We investigated reliability and robustness of proposed 264-gene signature using each of the expression data sets generated from different platforms.

3.3.3 Expression in Metastasis Tissues

Similarly, we carried out differential expression analysis for a subset of Shaffer et al. data set consist of 54 normal and 67 Mets samples and identified 1328 genes were significantly dysregulated. Among the total identified 1328 genes, 1310 have shown over-expression whereas 18 genes were under-expressed.

Table 4. Leave one out cross-validation classification of normal versus Mets signature (877-genes).

Author	Expression Set	No. of Samples	Accuracy (%)	Sensitivity (%)	Specificity (%)
Musella et al. (2013)	GSE37182	N=88, T=84	100	100	100
Shaffer et al. (2009)	GSE41258	N=54, T=186, M=67	93.39	96.82	100
Agesen et al. (2013)	GSE49355	N=18, T=20, M=19	93.33	93.17	96.6

Likewise, analysis using a subset of Musella et al. study consists of 18 normal and 19 metastatic samples have shown deregulation of 3122 genes, mostly 3098 showing over-expression (log FC > 1). We also focused our attention towards comparison of differentially expressed profiles among a tumour versus Mets tissues comparison and identified 124 genes were common between the resulted gene lists. We observed a number of notable genes were previously linked with metastases in colon cancer.

Table 5. Leave one out cross-validation classification of a tumour versus Mets signature (124-genes).

Author	Expression Set	No. of Samples	Accuracy (%)	Sensitivity (%)	Specificity (%)
Musella et al. (2013)	GSE37182	N=88, T=84	100	100	100
Shaffer et al. (2009)	GSE41258	N=54, T=186, M=67	72.96	72.82	95.53
Agesen et al. (2013)	GSE49355	N=18, T=20, M=19	76.86	80.39	96.6

3.3.4 124-gene metastatic signature identified patients associated with poor outcome in independent data set

An independent human colon cancer gene expression and clinical data set was used to test the ability of 124-gene signature that discriminate between patient associated with cancer reoccurrence, overall survival and disease-specific survival. 238 patients with histopathological properties of age, gender, ethnicity, stage and grade were available for analysis. We observed, patients with higher grade (grade 3) across all the grades in independent set has significantly better OS (p=0.001, HR=2.61(CI 1.43-4.79); p=0.16, HR=1.42(CI 0.86-2.35), respectively) and DSS (p=0.00, HR=2.41(CI 1.28-4.53); p=0.25, HR=1.35(CI 0.80-2.28) compare to low grade patients.

Similarly, we determine the relative risk of reoccurrence and cancer-related deaths. We observed a significant association of 124-gene signature with the risk of reoccurrence when analysed across all the tumour grades (p=0.0003, HR=1.74(CI 1.28-2.37)). We also analysed that the relative risk of reoccurrence has increased with the increase of tumour grade in patient samples (grade 3 (p=0.0005, HR=2.94(CI 1.59-5.46))) (Figure 3).

We performed prognostics analysis for the 124-genes signature derived from the comparison of tumour versus Mets tissues. For this purpose, we used independent data set (GSE17538) from the study conducted by Smith et al. (Smith et al., 2010) derived from metastatic colon cancer.

We tested 124-genes association with the clinical endpoints such as Overall survival (OS), Disease-specific survival (DSS) and Disease-free survival (DFS) across all the grades (grade 1, 2 and 3) by building Cox proportional hazard (PH)

model. We build classifiers using genes from signature genes and their corresponding values from the training set for the calculation of Wald score for each of the genes in the classifier. Log-rank tests P-Value were computed for both univariate and multivariate Cox model for OS, DSS and DFS. Similarly, Kaplan-Meier estimates were calculated for each endpoint.

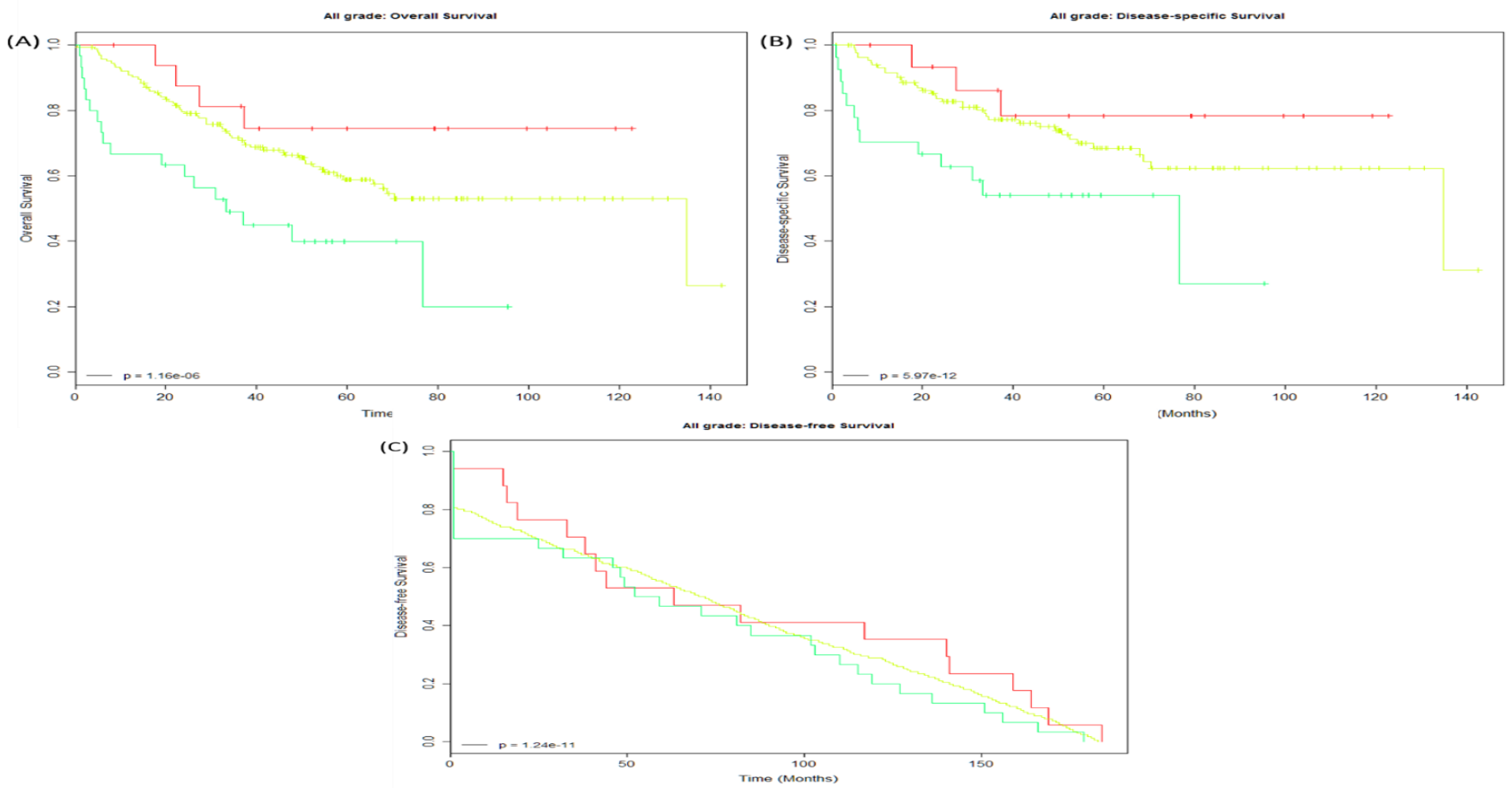


Figure 2. The 124-gene classifier as tested in the independent data set across all grades. Kaplan–Meier estimates of overall and disease-specific survival in the test set. Expression data for probes corresponding to the 124-gene recurrence classifier were used to build the Cox proportional hazard model from patient data in the Vanderbilt dataset. Plots represent survival analyses in the independent patient data set (A) Overall survival, (B) disease-specific survival analyses and (C) disease-free survival

3.3.5 The Cancer-focused Genes

We further analysed identified signatures by performing meta-analysis among inter-study signature derived from individual comparisons of normal versus tumour, normal versus Mets and tumour versus Mets tissues by comparing the similarities between them. We observed overlapping of 184 genes between normal versus tumour and normal versus Mets gene lists. However, 64 genes of normal versus tumour comparison represents a strict tumour-specific (those genes which are not significantly dysregulated in other comparison) pool for which the functional analysis identified their targeted pathways involved: cell cycle, acute myeloid leukemia, progesterone-mediated oocyte maturation, TGF-beta signaling pathway, role on ran in mitotic spindle regulation and G1-phase progression by my.

Similarly, in normal versus Mets and tumour versus Mets, a total of 701 genes were differentially expressed among met tissues. We tracked the significant pathways involved: immune response, lipid metabolism and PPAR signalling pathway.

3.4 Discussion

The purpose of the present study to find possible marker gene sets for colorectal cancer by using a two-step bioinformatics analytics. We performed meta-analysis using publicly available GEO expression profiles of normal, tumour and metastatic tissues for the discovery of robust signature involving pathogenesis of colon cancer.

We identified cross-study 267-gene signature from the comparison of normal versus primary tumour samples across all the data sets that may be vital for the

diagnosis of colon cancer. The functional analysis of 267-genes has revealed the involvement of cell cycle, cell-signaling and metabolic regulated pathways as reported in the previous studies (Moreno and Sanz-Pamplona, 2015, Planutis et al., 2014). We further tested the robustness of gene signature using cross-validation, which shows excellent 90.53% overall-average accuracy-rate across all the expression data sets. We also observed the Agesen et al. expression validated with higher error-rate compare too other two data sets in the validation cohort. A close examination of the cohort present possible to the explanation of these results; Agesen et al. study samples include stage IV tissues whereas Musella et al. and Shaffer et al. samples were derived from slightly earlier stages I-IV. However, we cannot rule out these variations are due to the difference in sample size and/or platform differences.

For the Mets tissues analysis, we identified two gene sets of deregulated genes from the comparison of normal, tumour and mets tissues. Further analysis of 124-genes deregulated among a tumour versus Mets tissues has shown involvement in key regulatory pathways such as complement cascade, the formation of the fibrin clot, extracellular matrix organization, collagen degradation and lipoprotein metabolism. Survival analysis of 124-gene signature using independent data sets has separated patients with high grades from lower grades when analysed for overall survival rate and disease-specific survival. The ability of 124-gene signature to discriminate between patient outcomes may be useful in patient prognosis, but further biological validation will be required. The prognostics results also show the positive correlation between the risk of reoccurrence and disease related deaths with the increase of tumour grade.

We also compared the similarities between the results of three signatures. The deregulated 64 genes were specific to normal versus primary tumour comparison have also shown significant linkages to cancer-related pathways. The other group of genes (701), strictly related to Mets tissues have also shown significantly involvement in pathways previously observed in colon cancer.

In conclusion, this study shows the importance of integrated techniques of individually conducted gene expression studies and provide further insights into understanding of colon cancer data for clinical purposes. The cross-validation analysis of gene signature shows samples scarcity and different platform used for generation of expression remains challenging area. This study has also shown valuable knowledge and future direction for the treatment of colon cancers but a more robust approach using multiple biological stage data may answer a question related to molecular heterogeneity.

CHAPTER 4 GENOME-WIDE MICRORNA AND MRNA INTEGRATED ANALYSIS OF COLON CANCER

4.1 Introduction

The majority of deaths related to colon cancer are due to metastases in the primary tumour lesions, and according to recent studies nearly half of the patients only survive for 5 years from the diagnosis of metastatic malignancy (Parkin et al., 2005). Various studies have highlighted the first site on the onset of metastatic colon cancer is regional lymph nodes and then its spread to the liver. Pathological studies on colon cancer cannot precisely predict patients with metastatic vulnerability to local lymph nodes and/or to distant organs.

The investigative studies on liver metastases have shown that it is often originated from colon cancers, and there have been such practical evidence where metastasis reading is the first and only finding in patients with an unknown primary tumour site (Pavlidis et al., 2003), and the discovery target sites can differentiate between primary hepatic lesions and liver metastasis from different possible origin sites can be therapeutic and prognostic value (Fernandez-Pineda et al., 2015). Therefore, there is an increasing urgency of novel diagnostic and prognostic biomarkers that could differentiate between the primary tumour and metastasis malignancies sites, as well as prediction of primary tumours having a tendency of metastasizing to other organs. Although, the association between colon cancer metastases and the mortality rate is very well studied the genomic mechanisms underlying tumour cell distribution and the primary tumour tendency for metastases is still poorly understood. However, the discovery of new class RNAs (miRNAs) with the regulatory role may be integral in malignant processes (Su et al., 2010, Glud et al., 2010).

miRNAs (microRNAs), are small (19–25 nucleotides) on coding RNAs, that have the ability to regulate mRNA genes expression by suppressing mRNA translation during post translational modifications, and/or causing mRNA degradation. miRNAs are known to be involved in important regulatory processes by targeting translation sites of multiple mRNA genes (Bartel, 2004), and are observed to expressed in tumour initiation and progression sites (Calin and Croce, 2006). c (Cho, 2007). The stimulating characteristics of miRNAs signatures being highly tissue-specific can be utilised to classify and investigate heterogeneous cancers and origin sites of colon metastases of unknown origin (Rosenfeld et al., 2008, Ramaswamy et al., 2001).

The aim of our study was to identify a microRNA signature that can differentiate primary and metastatic tumours, its stages and tumour grades in patients with colorectal carcinoma. Furthermore, we focused on the identification and evaluation of miRNAs potentially involved in prognosis and functional processes during metastatic progression. For this purpose, we performed differential expression analysis between histopathological groups and identified significantly dysregulated miRNAs followed by the target prediction analysis to establish their tumour transcriptional phonotypes.

4.2 Results

4.2.1 Subtype-specific miRNAs

In order to understand the impact of dysregulated miRNAs in the formation of a tumour transcriptional phenotypes, we investigated miRNA expression patterns across different histopathological tumour groups. We performed differential

expression analysis across tumour groups and identified approximately 70 highly dysregulated (p-value < 0.001) miRNAs (table 6).

Table 6. Top differentially expressed miRNAs among histopathological groups.

No.	Differentially expressed miRNAs	Gender	Mets	Grade	Adjuvant chemotherapy	Stages
1	hsa-miR-378*-4373024		✓	✓	✓	✓
2	hsa-miR-200c*-4395397					✓
3	hsa-miR-106b*-4395491		✓		✓	✓
4	hsa-let-7f-1*-4395528					✓
5	hsa-miR-15b*-4395284		✓		✓	✓
6	hsa-miR-424*-4395420		✓	✓		✓
7	hsa-miR-543-4395487					✓
8	hsa-miR-628-3p-4395545					✓
9	hsa-miR-769-5p-4395186		✓		✓	✓
10	hsa-miR-550-4395521			✓		✓
11	hsa-miR-99b*-4395307					✓
12	hsa-miR-26b*-4395555					✓
13	hsa-miR-155-4395459			✓		✓
14	hsa-miR-135a*-4395343		✓	✓		✓
15	hsa-miR-25-4373071					✓
16	hsa-miR-324-5p-4373052					✓
17	hsa-miR-27b-4373068		✓	✓		✓
18	hsa-miR-181c*-4395444					✓
19	hsa-miR-183*-4395381					✓
20	hsa-miR-335-4373045		✓	✓		✓
21	hsa-miR-26a-1*-4395554		✓	✓		✓
22	hsa-miR-143*-4395257		✓		✓	✓
23	hsa-miR-432-4373280				✓	✓
24	hsa-miR-16-4373121					✓
25	hsa-miR-101-4395364					✓

26	hsa-miR-505-4395200				✓	
27	hsa-miR-938-4395292				✓	
28	hsa-miR-193b-4395478				✓	
29	hsa-miR-15a*-4395530				✓	
30	hsa-miR-18a-4395533				✓	
31	hsa-miR-526b*-4395494				✓	
32	hsa-miR-376a-4373026				✓	
33	hsa-miR-770-5p-4395189				✓	
34	hsa-miR-10b-4395329				✓	
35	hsa-miR-92a-4395169			✓		
36	hsa-miR-146b-5p-4373178			✓		
37	hsa-miR-31-4395390			✓		
38	hsa-miR-19b-1*-4395536			✓		
39	hsa-miR-672-4395438			✓		
40	hsa-miR-875-3p-4395315			✓		
41	hsa-miR-551b-4380945			✓		
42	hsa-miR-149-4395366			✓		
43	hsa-miR-92a-1*-4395248			✓		
44	hsa-miR-549-4380921			✓		
45	hsa-miR-550*-4380954			✓		
46	hsa-miR-504-4395195			✓		
47	hsa-miR-142-3p-4373136			✓		
48	hsa-miR-330-5p-4395341			✓		
49	hsa-miR-194*-4395490			✓		
50	hsa-miR-148b*-4395271			✓		
51	hsa-miR-221-4373077			✓		
52	hsa-miR-194-4373106			✓		
53	hsa-miR-339-3p-4395295			✓		
54	hsa-miR-449a-4373207			✓		

55	hsa-miR-100*-4395253			✓		
56	hsa-miR-146b-3p-4395472			✓		
57	hsa-miR-378-4395354			✓		
58	hsa-miR-93*-4395250			✓		
59	hsa-miR-628-5p-4395544			✓		
60	hsa-miR-486-5p-4378096			✓		
61	hsa-miR-571-4381016			✓		
62	hsa-miR-335*-4395296		✓			
63	hsa-miR-323-3p-4395338		✓			
64	hsa-miR-497-4373222		✓			
65	hsa-miR-125a-3p-4395310		✓			
66	hsa-miR-655-4381015		✓			
67	hsa-miR-28-3p-4395557		✓			
68	hsa-miR-636-4395199		✓			
69	hsa-miR-923-4395264	✓				
70	hsa-miR-518b-4373246	✓				

We combine dysregulated miRNAs of all histopathological tumour groups in one table to differentiate between specific miRNAs which belong to the particular histopathological group. As we observed in above table there are 7miRNAs are which only exist in metastasis, 12 miRNAs in stages, 2 miRNAs in gender, 27 miRNAs in grade and 9 miRNAs in Adjuvant chemotherapy. These miRNAs are specific only to histopathological groups.

4.2.2 miRNAs differentiate primary and metastatic colon tissues

In order to investigate miRNAs profile expression among the tissues classes, we performed differential expression analysis among the 47 primary and 18 metastatic colon tissues and identified 17 miRNAs dysregulated among the patients, some of them have been reportedly associated with tumour activities in

colon cancer in the past. We observed up-regulation of 6 miRNAs (hsa-miR-636, logFC 1.57; hsa-miR-655, logFC 1.33; hsa-miR-135a*, logFC 1.28; hsa-miR-26a, logFC 1.25; hsa-miR-335*, logFC 0.89; hsa-miR-335, log FC 0.83) in primary tissue. Previously, hsa-miR-636 dysregulation has been reported in association with survival (Slattery et al., 2016) and down-regulation of hsa-miR-135a* linked to cell cycle regulation in colon adenoma cells (Schlormann et al., 2015). Similarly, up-regulation of hsa-miR-26a is associated with down regulation of CDK6 mRNA and induce apoptosis of colon cancer cells (Konishi et al., 2015). We observed up-regulation of two members of miR-335 family, previously involved in process of multiple tumorigeneses in colon tumours (Wang et al., 2010).

We also identified 11 up-regulated miRNAs (hsa-miR-769, hsa-miR-28, hsa-miR-27b, hsa-miR-125a, hsa-miR-497, hsa-miR-424*, hsa-miR-378*, hsa-miR-323, hsa-miR-106b*, hsa-miR-15b*, and hsa-miR-143*) in metastatic tumors as compared to primary colon. Previously, down-regulation of hsa-miR-28 (Almeida et al., 2012) has been reportedly linked to reduced cell proliferation, migration and invasion in vitro, so the over-expression of hsa-miR-28 in our study indicate oncogenic effects on processes like cell proliferation and migration. Further studies based on dysregulation of hsa-miR-28 in colon cancers may provide reciprocate details of genes likely to involved in metastases. Higher expression of hsa-miR-27b has been associated with proof clinical response (Rasmussen et al., 2013), reportedly involved in suppressed tumour growth, cell adhesion, and invasion (Matsuyama et al., 2016). miRNAs hsa-miR-424* and hsa-miR-378* up-regulation reported being involved in lymph node metastases and poor prognosis (Wang et al., 2012, Wang et al., 2010). Other miRNAs with oncogenic

activities are hsa-miR-125a inhibit cell proliferation and induce apoptosis by targeting BCL2, BCL2L12 and Mcl-1 (Tong et al., 2015) and miR-106b mediate inhibition of LT97 cell proliferation (Schlormann et al., 2015).

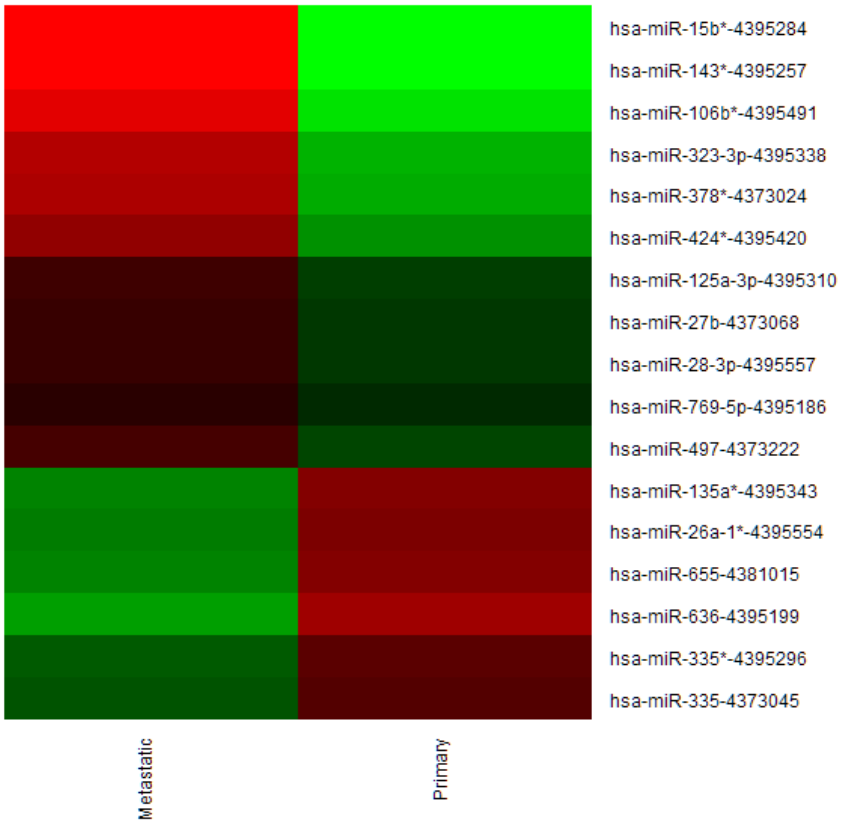


Figure 3. Heatmap representation of miRNAs differentially expressed among primary and metastatic tissue classes.

We identified 11 up-regulated miRNAs (hsa-miR-769, hsa-miR-28, hsa-miR 27b, hsa-miR-125a, hsa-miR-497, hsa-miR-424*, hsa-miR-378*, hsa-miR-323, hsa-miR-106b*, hsa-miR-15b*, and hsa-miR-143*) in metastatic tumors as compared to primary colon. Previously, down-regulation of hsa-miR-28 (Almeida et al., 2012) has been reportedly linked to reduced cell proliferation, migration and invasion in vitro, so the over-expression of hsa-miR-28 in our study indicate oncogenic effects on processes like cell proliferation and migration.

4.2.3 miRNAs differentially expressed among the stages

Studies based on miRNAs specific to tumour stage and survival are vital for the understanding of tumour progression and origin sites (Slattery et al., 2015a). Therefore, we investigated miRNA expression across the different stages of colon cancer by performing univariate ANOVA analysis and identified 25 miRNAs differentially expressed (p-value 0.001) among the four stages. Of these, 11 showed over-expression in stage I, 15 in stage II, 15 in stage III and 14 miRNAs in stage IV.

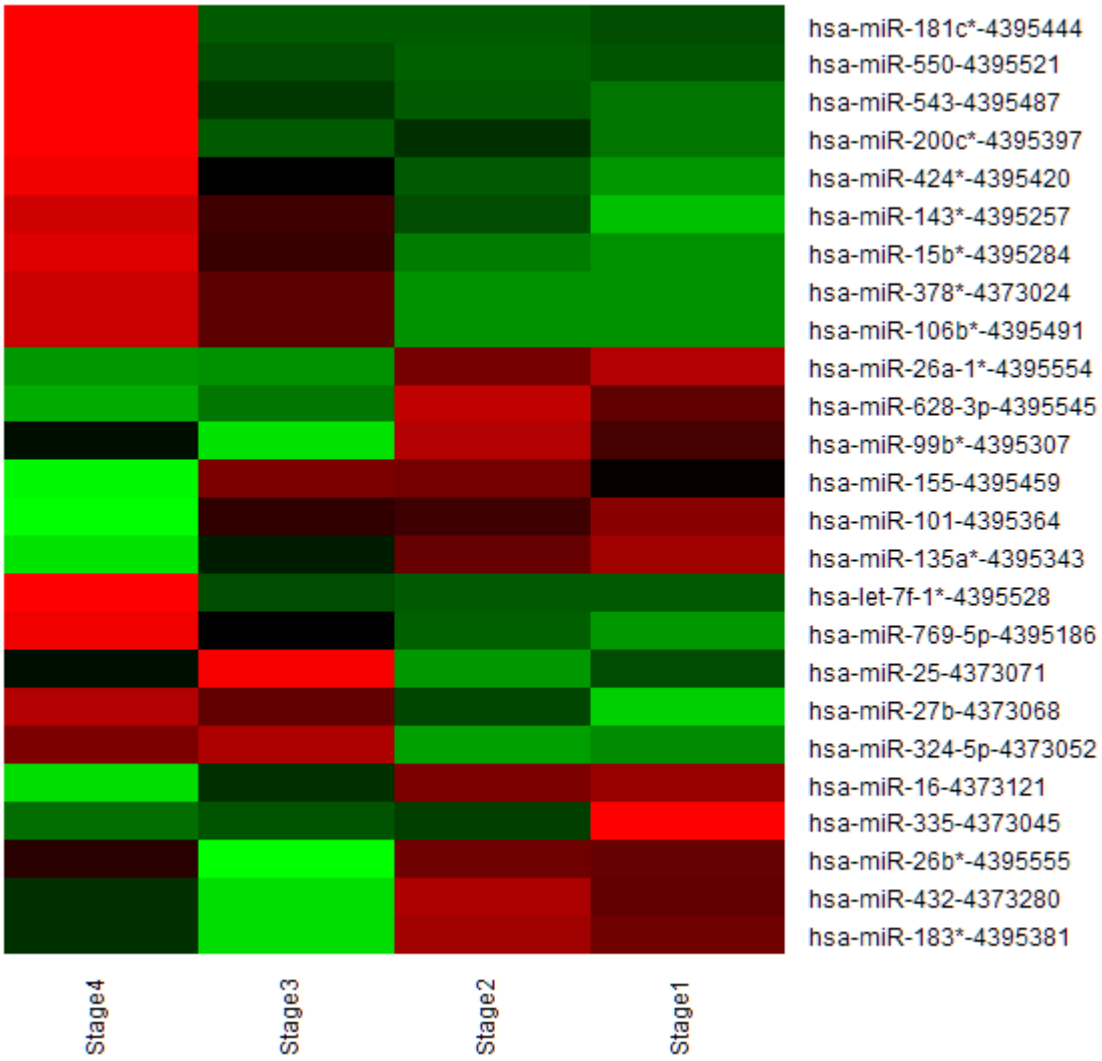


Figure 4. Heatmap representation of miRNAs differentially expressed among four stages of colon cancers.

Surprisingly, none of the miRNA either showed continuous over-expression or under-expression from stage I to stage IV during the analysis which provides strong evidence of miRNAs ability of stage specificity. We observed similarities in the expression changes of miRNAs between stage I and II, and nearly the same numbers of miRNAs altered its expression from stage III to stage IV (heatmap). We also observed significantly decreased expression of miRNAs compare to stage III and IV in stage I and II tumours, suggests a shift in the pattern changes of miRNAs and can be differentiated in the presence of this signature. The overall

behaviour of miRNAs between primary stage I and II was same but upon evaluating individual miRNAs we observed a significant increase in expression of two miRNAs in stage II colon tumours (hsa-miR-200c*, logFC1.39 and hsa-miR-181c*, log FC 1.06). Further analysis of these miRNAs could only reveal their ability to distinct between the stage I and II tumours along with their vulnerability to undergo metastases.

Another interesting theme observed during analysis was that none of the miRNA expressed in primary stage I and II was associated with metastases, however, five significantly miRNAs (hsa-miR-15b*, hsa-miR-143*, hsa-miR-106b*, hsa-miR-378* and hsa-miR-424*) over-expressed in stage III and IV. We further evaluated the relationship of these miRNAs for association with survival and their anti-correlated impact on transcription sites of mRNA.

4.2.4 Histological grades

The investigation on expression patterns across the tumour grades shows significant (p-value 0.01) dysregulation of 34 miRNAs from well, moderate and poorly differentiations of colon tumour cells. Of these, 14 miRNAs showed over-expression in well, 17 in moderate, and 21 in poorly differentiated tumour grades. As expected, we observed decreasing expression from well to poorly differentiated miRNAs.

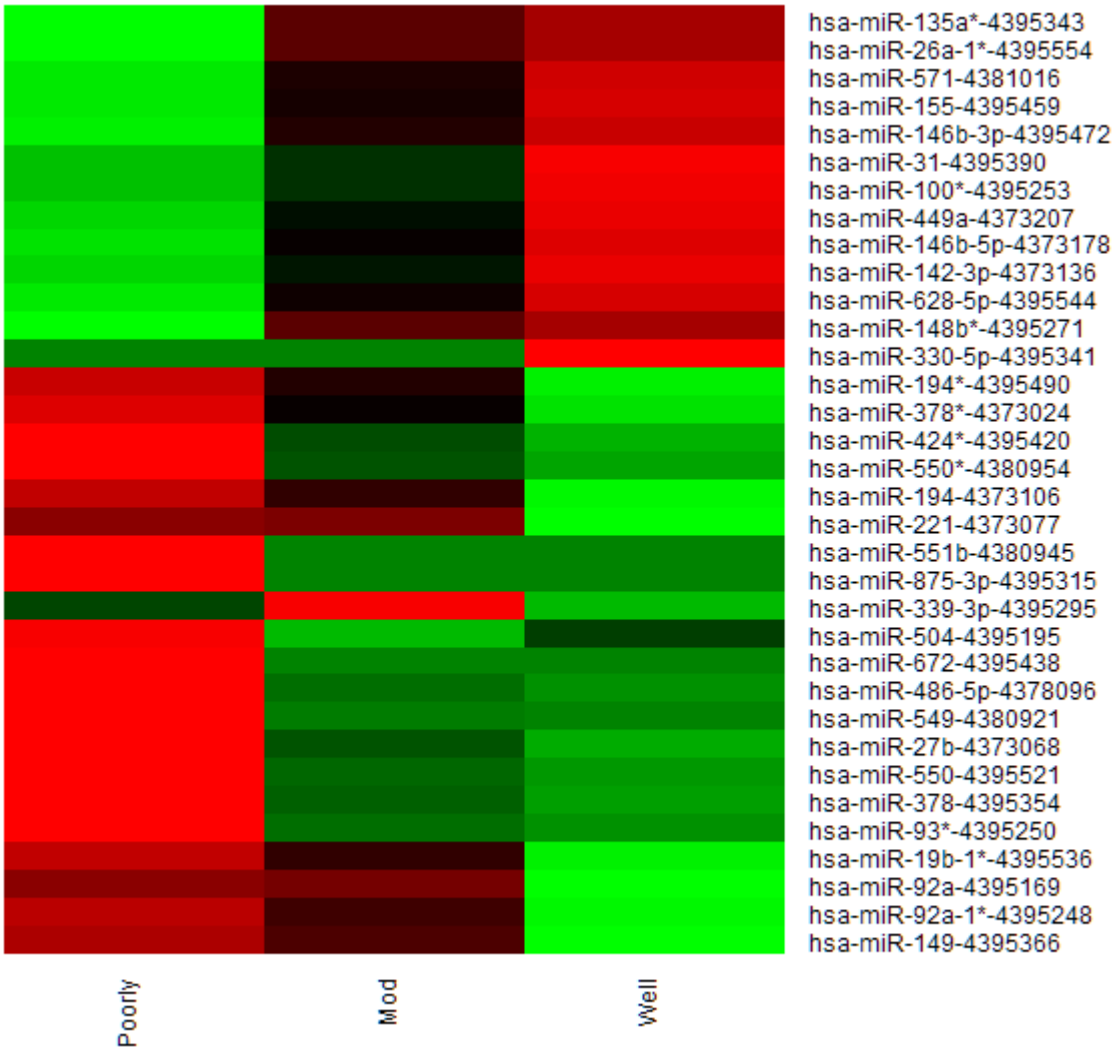


Figure 5. Heatmap representation of miRNAs differentially expressed among three grades of colon cancer.

We have also investigated the association of tumour grades with primary and metastases colon tumour and observed nearly three miRNAs (hsa-miR-378*, hsa-miR-424*, and hsa-miR-27b) have shown higher expression in poorly tumour differentiation as well as in metastatic colon tissues. However, we observed over-expression of two miRNAs (has-miR-135a*, log FC 2.90 and hsa-miR-26a, log FC 2.94) in well-differentiated tumour grades as well as also shown higher expression in primary tumour colons.

4.2.5 Adjuvant Chemotherapy

Determining subset of high risks patient likely to get the benefit of chemotherapy could add valuable information to the clinical features in colon cancer metastasis potential and drug resistant (Li et al., 2016). Genetic therapies treatments with miRNAs in a combination of chemotherapy and surgeries are essential for suppression of tumour growth in advanced-stage colon cancers (Okamoto et al., 2016). Here we report dysregulation of 15 miRNAs after comparison of groups of patient served with adjuvant chemotherapy against comparing to who hasn't. Among the 8 up-regulated miRNAs are (hsa-miR-15b*, logFC 2.59; hsa-miR-143*, logFC 2.13; hsa-miR-106b*, logFC 1.92; hsa-miR-505, logFC 1.65; hsa-miR-378*, logFC 1.36; hsa-miR-18a, logFC 0.75; hsa-miR-10b, 0.60; hsa-miR-769, logFC 0.49). A closer look at the results shows that nearly five miRNAs have responded with high expression in patients served with chemotherapy also showed over-expression in metastatic tissues. Similarly, they have also shown correlation with at stage III and IV colon tumours.

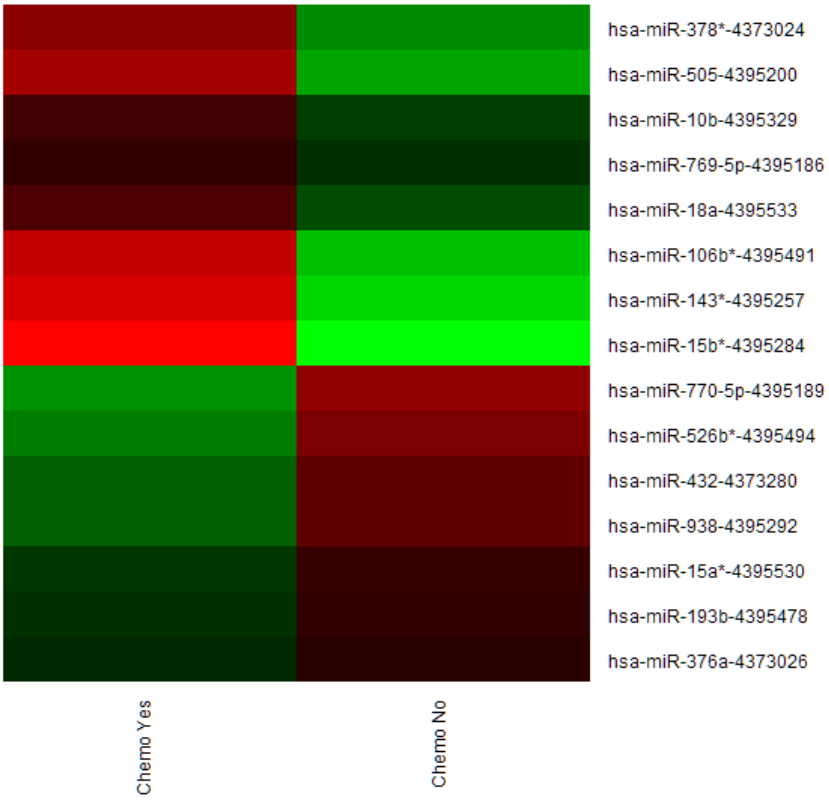


Figure 6. Heatmap representation of miRNAs differentially expressed among histological groups of colon cancer patient who were treated with adjuvant chemotherapy versus who received no treatment.

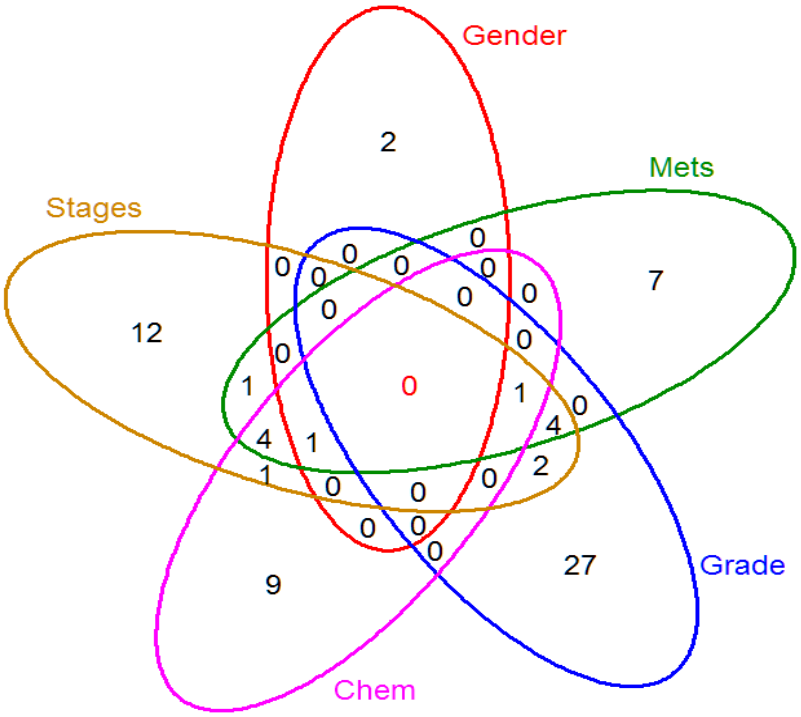


Figure 7. Venn diagram summarising miRNAs dysregulated among five histological groups.

We combine dysregulated miRNAs of all histopathological tumour groups on Venn diagram to differentiate between specific miRNAs which belong to the particular histopathological group. As we observed in above diagram there are 7miRNAs are which only exist in metastasis, 12 miRNAs in stages, 2 miRNAs in gender, 27 miRNAs in grade and 9 miRNAs in Adjuvant chemotherapy. These miRNAs are specific only to histopathological groups.

4.2.6 miRNA associated with prognosis in colon cancers

We performed Cox-regression univariate analysis in order to identify miRNAs whose expression is associated with clinical outcome. Analysis was for conducted using all the colon tumour samples with respect to overall survival (OS) and disease-free survival (DFS) from the point of diagnosis till clinical end points

(death or recurrence of disease). As a result, we identified 10 miRNAs associated with clinical outcomes (table 7). Among these, 9 of them were significantly associated with OS and only one miRNA was linked to DFS. We further analyse these prognostic miRNAs by assessing their expression in different histopathological groups and found almost all the prognostic miRNAs were significantly dysregulated between them. For example, prognostic miRNAs (hsa-miR-378* and hsa-miR-15b*) were dysregulated in metastatic patients, three grades (poorly, moderate and well differentiated), adjuvant chemotherapy (yes or no), and among the four stages of tumour progression. Similarly, we identified individual prognostic miRNAs which have shown dysregulation in one particular histology comparison such as, has-miR-183* uniquely expressed among the four stages of tumour samples, and two miRNAs (hsa-miR-92a-1* and hsa-miR-330-5p) when we performed comparisons among the patient with different tumour grades.

Table 7. Summary table of miRNAs associated with survival outcomes.

	miRNAs	**miRNAs significantly associated with prognosis**			
No.		**HR**	**Lower**	**Higher**	**P-value**
1	hsa-miR-378*-4373024	0.81	0.66	0.99	3.71E-02
2	hsa-miR-15b*-4395284	0.81	0.69	0.95	5.56E-03
3	hsa-miR-628-3p-4395545	1.3	1.04	1.62	1.95E-02
4	hsa-miR-135a*-4395343	1.45	1.06	2	1.85E-02
5	hsa-miR-183*-4395381	1.64	1.22	2.22	1.51E-03
6	hsa-miR-330-5p-4395341	6.02	2.49	14.56	2.30E-11
7	hsa-miR-323-3p-4395338	0.83	0.69	1	4.19E-02
8	hsa-miR-125a-3p-4395310	0.56	0.33	0.97	3.69E-02
9	hsa-miR-655-4381015	1.36	1.08	1.73	8.05E-03

10	hsa-miR-92a-1*-4395248	0.73	0.54	1	4.10E-02

All the prognostic miRNAs were further divided into risk groups (high or low) according to their risk-score predictor. Prognostic miRNAs exhibiting Hazard Ratios less than one (HR < 1) were defined as "Protective" and miRNAs associated with Hazard Ratio greater than one (HR > 1) as "Risk-associated". Five of the miRNAs associated with better prognosis (hsa-miR-378* (HR = 0.81, CI = 0.66-0.99), hsa-miR-15b* (HR = 0.81, CI = 0.69-0.95), hsa-miR-323-3p (HR = 0.83, CI = 0.69-1), hsa-miR-125a-3p (HR = 0.56, CI = 0.33-0.97), and hsa-miR-92a-1* (HR = 0.73, CI = 0.54-1). Similarly, prognostic miRNAs such as hsa-miR-628-3p (HR = 1.3, CI = 1.04-1.62), hsa-miR-135a* (HR = 1.45, CI = 1.06-2), hsa-miR-183* (HR = 1.64, CI = 1.22- 2.22), hsa-miR-330-5p (HR = 6.02, CI = 2.49-14.56), and hsa-miR-655 (HR = 1.36, CI = 1.08-1.73) were linked to worse prognosis. All of the risk-associated miRNAs were absent when we compared patients who have received adjuvant chemotherapy versus who hasn't. A closer look at risk-associated miRNAs shows that high-risk hsa-miR-183* has shown higher expression in stage I and II compared to stage III and IV, could be used as a biomarker for early detection of colon cancer. We further performed Kaplan-Meier analysis using prognostic miRNAs in order to calculate and illustrate survival curves.

Further analyses were performed on prognostic factors by conducting univariate and multivariate analysis using histopathological information on its own (without the expression data). As a result, we found histological factor stage is significantly associated with prognosis of DFS (table 8) whereas histological factors such as stage, tissue type and use of adjuvant chemotherapy significantly associated with

OS prognosis (table 9). Following the cox regression analyses, we then assessed the quality of fitted models using analysis of deviance (-2 log likelihood) for the selection of co-variate which could impact on the association of prognostic factors with miRNA expression on the outcome prediction. The deviance-score analysis shows one-factors (stage level) model can be a best-suited model for DFS prognosis whereas, prognostic factors tissue type and stage level models for OS prognosis. Almost all the miRNAs exhibited their ability as an independent prognostic factors when evaluated using multivariate models for DFS and OS.

Table 8. Summary table of prognostic factors associated with DFS.

Disease-free survival								
Histopathological factors			Univariate			Multivariate		
		N(n)	HR	%95 CI	P-Value	HR	%95 CI	P-Value
Gender	Female	25	1			1		
	Male	40	0.9686	0.2591 3.621	0.962	0.9704	0.25327-3.718	0.965
Tissue Type	Tumor	47	1			1		
	Mets	18	2.893	0.5843-14.33	0.193	0.4891	0.03775-6.337	0.584
Tumour Grade	Continuous	65	1.032	0.3754-2.834	0.952	1.0775	0.35709-3.251	0.895
Stage	Continuous	65	2.464	1.049-5.786	0.03845	3.251	0.64426-16.405	0.153
Adjuvant Chemotherapy	No	27	1			1		
	Yes	38	2.558	0.5244-12.48	0.2454	1.2761	0.22312-7.298	0.784

Further analyses were performed on prognostic factors by conducting univariate and multivariate analysis using histopathological information on its own (without the expression data). As results, we found histological factor stage is significantly associated with prognosis of DFS in the table above.

Table 9. Summary table of prognostic factors associated with OS.

Overall survival								
Histopathological factor			Univariate			Multivariate		
		N(n)	HR	%95 CI	P-Value	HR	%95 CI	P-Value
Gender	Female	25	1			1		
	Male	40	1.58	0.682-3.687	0.28	1.5129	0.63151-3.6244	0.35299
Tissue Type	Tumor	47	1			1		
	Mets	18	5.059	2.243-11.41	0.0000938	1.2636	0.25292-6.3128	0.84389
Tumour Grade	Continuous	65	0.6256	0.2618-1.495	0.291	0.9112	0.36127-2.2984	0.77561
Stage	Continuous	65	2.758	1.646-4.623	0.0001174	3.4753	0.64426-16.405	0.0154
Adjuvant Chemotherapy	No	27	1			1		
	Yes	38	0.72	0.3285-1.619	0.43	0.2201	1.26872-9.5193	0.001

Histological factors such as stage, tissue type and use of adjuvant chemotherapy significantly associated with OS prognosis (table 9). Following the cox regression analyses, we then assessed the quality of fitted models using analysis of deviance (-2 log likelihood) for the selection of co-variate which could impact on the association of prognostic factors with miRNA expression on the outcome prediction.

4.2.7 miRNA, mRNA coupling analysis

In order to established relation between miRNAs and their respective mRNA gene targets, we carried out target prediction analysis followed by Pearson correlation analysis. The main purpose of this proposed model is the biology of miRNA and the methods in prediction of gene targets. It has been widely known that the binding of miRNA during transcriptional activity degrade predicted targets, so only the anti-correlated predicted targets can prove to be the real ones. Second, the predicted targets are usually unreliable and cannot be implied in biological observation. Thirdly, only miRNAs showing the anti-correlation between its targets can and are more likely to play a role in functional activities.

So for each prognostic miRNA, a list of putative candidate genes was extracted using five different predicting algorithms (table 10) followed by an independent correlation analysis. We focused on only anti-correlated miRNA, mRNA pairs according to the biology of miRNAs. The independent correlation analysis was performed among the pairs due to lack of correlation between the miRNAs and the predicted targets obtained from the five algorithms. For miRNA:mRNA pairs anti-correlation analysis, we isolated each miRNAs targets predicted by the five prediction algorithms and then extracted expression of the matching target genes from the mRNA expression sets. In total, we observed over-lapping of 1831 (figure 9) genes among the predicted targets and the mRNAs genes differentially expressed among the colon subtypes. The top-anti correlated genes are highlighted in the table (table 11).

Table 10. Predicted targets of prognostic miRNAs.

hsa-miR-655	APC, CDKN1C, DMD, FBN1, HK1, SGCB, MTM1, MYO5A, PTCH1, PAX6, EGR2, GPD2, APP, IFNG, LTBP1, HMGCR, IMPDH1, NR3C2, PPP3CA, ATP2B1, ATP2B4, SMAD5, PTER, GALK2, ZNF470, SIAH1, BRWD1, EI24, TULP4, ZDHHC9, CGGBP1, RBM33, TMEM64, CXCR4, EML1, FNIP1, MLLT10, ZCCHC11, PDLIM5, SEH1L, RFFL, SP3, GOPC, GOLGA8B, TNRC6B, TGFBR2, FNBP1L, CD47, NDEL1, ARFIP1, AHDC1, ARHGAP5, RHOBTB1, PPM1B, PHF12, CYFIP2, PDE4B, AGPAT3, USP9X, CDC42, PCDH17, WHSC1, SHPRH, HIPK3, ZNF436, WIPF1, KTN1, PAICS, VAV3, USP6NL, STRN3, CTBP2, TCF4, MEX3A, ACLY, XPO7, PHACTR2, SENP6, KDELR2, ZFAND5, RGS4, PCDH19, ZFYVE16, UBE4B, FBXO45, ADAM10, ADD3, ADM, ANK2, SLC25A5, ARHGAP6, CALU, CHD2, CHUK, CREBL2, DYRK1A, EP300, EPAS1, ESRRG, FOXC1, ING2, ATP2A2, BCL6, KLF5, CBFB, CDH2, CEBPG, COL4A1, COL11A1, GADD45A, DR1, EGR1, EIF4A2, EIF4E, GAS1, GNAI1, GNAQ, HOXB3, INHBB, ITGB1, ITGB8, KDR, MARCKS, MBD1, MEIS1, PPP1R12A, PDGFA, PIK3C2A, PLN, PNN, PPP1CB, PRKD1, MAPK1, MAPK9, PTPN1, PTX3, RAB5B, RDX, RGS1, SATB1, MAP2K4, SNAI2, SMARCE1, SNRPE, SNX1, SSB, RPS6KB1, UVRAG, VSNL1, WEE1, XPO1, YWHAZ, ZNF91, PTP4A1, FZD5, DYRK2, MAP4K3, GAS7, BCAS1, AGPS, CASK, TMEFF1, SLC25A12, PPAP2B, HSD17B6, NCOA1, EIF4G3, DYRK4, TRIM24, SOCS2, EIF2S2, SEMA5A, BTAF1, GMFB, RAB5A, PTTG1, SOCS6, JMJD1C, HOMER1, ACTL6A, BNIP2, DACH1, EPHA4, ETV5, ACSL3, ACSL4, BPTF, MLLT3, TSPAN7, CDK2AP1, ETF1, CCPG1, KIF3B, MAP4K4, ROCK2, SCAMP1, VPS4B, TMEM59, SPAG7, SEC22B, ETV1, JARID2, DNAJA3, CEBPB, GTF2H1, NFIL3, PPP1R3C, VEGFC, MAFB, RANBP9, LRRC32, KIF20A, YAF2, TOB1, SRRM1, SPRY1, IRX5, SMNDC1, MEOX2, NAB1, UCHL3, CD164, SEMA3A, SPON1, NEDD4, PBX3, GNAI3, EXOC5, DCTN6, CHL1, HSPH1, HOXA2, LMO4, CBX1, PNRC1, CALM1, RAC1, RBBP6, ZNF22, NR4A3, TLE4, WWP1, UBL3, WBP4, SEC63, RNF13, PAXIP1, CCT5, ZFPM2, MACF1, BACE1, CHIC2, CD2AP, FZD4, NT5C2, RAB3GAP1, RYBP, HEY2, KPNA6, MYO10, RAB3GAP2, SETDB1, ZNF281, BRD7, FLRT3, GREM1, AK2, BAZ2B, CLIC4, NRG1, SSX2IP, CHMP2B, TRPS1, SETD2, AP2A1, ZNF318, RNF11, DAPP1, SESN1, GOLIM4, NUP93, SERTAD2, EPM2AIP1, CEP350, UBAP2L, RNF44, WDR47, STK38L, MON2, WAPAL, ZNF423, TBC1D9, ZCCHC14, ANKRD28, LPHN3, AGTPBP1, SATB2, TRIM2, SASH1, DOCK9, LARP1, ADNP, ZNF521, KBTBD2, OPA1, WSB1, COL5A3, TRIM33, PHF20L1, UBE2J1, HECA, PACSIN3, BIRC6, FAM8A1, TUBE1, KLF2, TOB2, HSPA14, COPS7A, CTDSPL2, PHF20, KLF3, NKRF, RNF111, CDKN2AIP, FAM46A, ANKHD1, WHSC1L1, YTHDF1, Mar-01, ECT2, TMEM30A, ASXL2, RSBN1, MBNL3, C17orf85, YOD1, ETNK1, ERRFI1, SLC38A2, ANKIB1, PLEKHA5, WDR44, ING3, NDFIP2, SLC39A10, EIF5A2, SMEK2, PELI1, USP31, HEG1, HACE1, WDFY1, KIAA1468, SH3RF1, LRCH2, CNTN3, GRAMD1A, NR2F2, RTN1, RAP2C, FAM60A, PELI2, GPBP1L1, TRIB2, SEC24A, XYLT1, RNF38, SLC30A5, REEP1, NADK, ATP13A3, FAM118B, SMC6, E2F8, PHC3, PGAP1, CEP135, FBXO11, NDFIP1, DICER1, DDHD1, SETD7, SGPP1, ANP32E, RNF146, COG3, CRISPLD1, ITCH, B3GNT5, FBXO30, TMEM117, KIAA1804, PCDH7, KBTBD8, FOXP1, GPR124, PURB, TP53INP1, TANC1, ABCC10, MAL2, CCND1, CCNE2, PHACTR3, DCBLD2, NIPBL, STXBP5, WDR20, FAM76B, ZNF23, PTPDC1, TMTC2, ARID2, UBE2E2, YTHDF3, PDIK1L, KBTBD6, AEBP2, USP43, CPNE8, NAP1L5, RDH10, RC3H1, AQP11, SLC25A26, VKORC1L1, PAN3, SESTD1, MTDH, CMTM4, PIK3R1, CPEB2, SRPK2, CXorf23, PRICKLE2, SS18L1, KCTD1, QKI, IL18RAP, ALMS1, FANCL, SPTBN1, STON1, CEBPZ, ODC1, LSM14B, ZBP1, EYA2, SEMG1, SGK2, TGM2, RPN2
hsa-miR-	IGF2BP2, EXOC7, SHANK3, KIF2A, RGS4, E2F3, GCG, MBD1, PPP1CB, MAP2K1, SMARCA2, DACH1, RPS6KA3, TSC22D1, MAB21L2, MAN1A2, PKIA, RNF139, BRCA1, RYBP, EPB41L3, TRPS1, KLHL20, NLK, HP1BP3, BNC2, NDE1, CHFR, ETNK1, NUFIP2, CFL2, NDFIP1, ZNRF3, C1orf198, TMEM129, GPX1, IER5L, SMAD2, SNX21, IRF2BP2, GNAS, ADD3, ANK2, BCAT2, LDB2, DPYSL2, FOXC1, CAV1, CTGF, GNAI2, GALNT4, RNGTT, FAP, MAP1B, PGRMC2, TRIO, BACE1, RAB3GAP1, G3BP2, NPTN, SEC14L2, CNOT4, BAZ2A, NRG1, TNFRSF21, SLK, SIPA1L1, RAB4B, EVL, CTDSPL2, F11R, FAM46A, Sep-11, GATAD2B, SCUBE2, BTF3L4, NR4A3, GANAB, DHX57
hsa-miR-323-3p	GJA1, HK2, MET, ATXN1, APP, ATRX, GAD1, ERBB2IP, IFRD1, BRWD1, CGGBP1, ENAH, MXI1, FNIP1, MLLT10, PPM1B, ARL5A, USP9X, ACTR3B, GLIS3, ZNF706, HIPK3, SENP7, PCBP2, TGFA, RAP1GDS1, TLE3, NRXN3, ACVR2A, KLF5, CCND2, CLCN4, PPP1CB, PRKAR1A, PTPRF, WEE1, AGPS, TNFSF11, CDKN1B, COL12A1, CREB1, DDX6, BPTF, DAPK1, EFNA3, FOS, TAF12, YAF2, MAN1A1, MAT2A, MAP3K5, NFYB, PBX3, PKN2, PGRMC2, MAP4K5, LRBA, WWP1, TMF1, ZIC2, ZFPM2, APPL1, FBXL5, G3BP2, MYLIP, FLRT3, LRP12, RNF11, TMOD3, PHF14, EDEM1, SERTAD2, DNAJC6, MELK, AAK1, USP33, ATP11A, LARP1, NIPBL, CNOT6, ZZZ3, UTP11L, GOLT1B, NLK, CDKN2AIP, FAM46A, SUV420H1, ANKHD1, COMMD8, PRPF40A, SCYL2, IFT57, TMEM30A, RSBN1, UBAP2, YOD1, PLEKHA5, C11orf30, SMEK2, GALNT1, GATAD2B, SRGAP1, PTBP2, HSPA2, MARCKSL1, FRY, TBL1XR1, PHC3, PGAP1, CHD9, CPEB4, SETD7, PCDH7, DAB2IP, PURB, SLC44A1, SLC30A7, KCTD12, UBLCP1, MIER3, UHRF2, JAZF1, ZNF326, CREB3L2, C1orf52, NOTCH2NL, TIMP3, SMAD5, ATP1B1, RUNX1, MTUS1, SMAD2, OSBPL8, TFEC, GOLGA8B, ARFIP1, PHF12, PDE4B, STAG2, TCF4, PHACTR2, PDS5A, PDE4D, ACVR1, ANK2, CDC5L, CREBL2, ALCAM, ARF6, CPE, FMR1, GCLM, LAMC1, LBR, PIK3C2A, PPP1CC, PSMD10, SH3BGRL, SPTBN1, VSNL1, RNMT, EGR3, MED1, MAP4K4, ATF1, ROCK1, DNAJB6, ARL4C, RBM5, PSME3, TBX3, NR4A2, PLCL1, RBBP6, FBXO8, RYBP, TRPS1, MTCH2, GHITM, NEK6, ZNF516, KIAA0355, PJA2, TBC1D4, WAPAL, SMG1, CAMTA1, LPHN3, TMEM87A, TIPARP, AUTS2, MTMR2, AKAP11, RAPGEFL1, PHF20, BNC2, KIAA1598, CAND1, SLC38A2, RIN2, ANKRD50, ABHD6, CGN, LRRN1, RAB18, ZNF148, XYLT1, BCL11A, GPBP1, RNF128, SGPP1, SNX27, ZNF566, PPTC7, TMED6, LSM14B, ZNF564, RBM24, ARL5B, SPTY2D1, B3GALTL, NUP43

hsa-miR-330-	ENG, ATXN1, GPD2, PAFAH1B1, RARG, ZDHHC9, BCAP29, BTBD11, FNDC3A, SHANK3, TCF4, XPO7, ITGA5, STC1, FZD5, TAGLN2, HRK, NR5A2, EGR3, EPHB3, ESRRA, ILF3, SNRPA, SLC27A4, TLN1, SOX12, SLC19A2, THBS3, BTBD3, METAP1, RCOR1, FAM53C, RFWD3, PAG1, GRIPAP1, LYRM2, MPP5, DGCR14, COPS7B, KCTD15, TBL1XR1, CBLL1, PVRL4, FAM107B, FBXL20, MAG, CHKB, MIER3, RALGPS2, ANKRD52, CMTM8, AUP1, ERBB2IP, C20orf194, VLDLR, NDEL1, CSNK1G3, SEC14L1, C4orf19, BCL2L1, DPYSL2, RHOA, CENPB, ITGA2, MAP1A, FGF18, DAG1, RPS6KA3, PRKAB2, MAP1B, MEOX2, PBX3, PKIA, SEC63, NRG1, VGLL4, CEP350, LARP1, YBX2, NLK, KLF3, F11R, PPP3CB, BCL11B, LCOR, GLIS2, C9orf24, BTN2A1, CAPN12, NHS
hsa-miR-183	NR3C1, CYP2B6, GOLGA7, PKP4, LPHN1, TTC7B, NCDN, KIAA0101, CSNK1G3, FRMD6, MIER1, KIAA0368, REPS2, TCF4, KIF2A, MRVI1, GNG4, PDE4D, ARHGAP6, ACVR2A, CLCN3, HBEGF, EGR1, LRP6, PLAGL2, PRKACB, PTPN4, RCN2, ROBO2, SLC6A6, STC1, CDK5R1, GMFB, HOMER1, ACVR1B, EPHA4, FOXA1, BUB3, ICA1, GTF2H1, RALA, YAF2, LHFPL2, CD164, SLC35A1, RGS14, MAP4K5, QKI, RAB35, DUSP10, XPOT, ZFPM2, AP3M1, RYBP, LIMD1, TRAM1, SESN1, BZW1, TOMM70A, ZHX2, STK38L, FRYL, ZDHHC17, ZFYVE26, CNOT6, RNF138, RAB8B, DPP8, PHF10, ING3, PLEKHA3, NUDT4, SCYL3, ARHGAP21, GPAM, NTN4, BACH2, ZDHHC6, WHSC1L1, TCF7L2, ARHGAP18, CYYR1, SLC44A1, FAM91A1, VPS37A, PLCB4, RORC, OSBPL8, NCK2, ERBB2IP, L3MBTL3, EI24, ZDHHC9, ENAH, TMPO, RHOBTB1, AMD1, SIRPA, TTC14, SLC25A36, CHD2, CTGF, HLF, IDH2, ITGB1, PFN2, PIM1, PLAG1, PPP2R2A, PRKCI, CX3CL1, SNX1, TCF12, UBE2V2, DCHS1, SOCS6, BNIP3L, RPS6KA3, MTMR6, MTA1, MED1, POLR2D, SEL1L, DMXL1, IRS1, MAP3K4, UNC13B, NFAT5, TLE4, MYO1B, SACS, ATP2C1, PDCD4, ZNF592, KIAA0355, SIN3A, AUTS2, YPEL5, LRRC1, AGPAT5, SMPD3, PLEKHA5, EML4, ANKRD50, GATAD2B, MBNL1, TSPYL4, GREM2, ANKRD13C, BRMS1L, PCGF5, FOXP1, MAL2, COLEC12, PPP1R14B, PRICKLE2
hsa-miR-135a	APC, GHR, GJA1, INSR, NPC1, COL5A2, ADRA2A, FKBP1A, RARA, RARB, SMAD2, EDA, SIAH1, SDCBP, IGF2BP2, CREB5, C6orf120, ARNTL, CSNK1G3, EXTL2, PPM1B, ATP6V1C2, PTPRD, ABCE1, CEP170, RRBP1, NET1, HIPK3, MEGF9, FOXN3, Sep-08, XPO7, ZFAND5, FBXO45, SLC25A5, TRIM23, CENPB, MEIS2, TNFRSF11B, P2RY2, PLCG1, PPP1CC, SMARCA2, SMARCE1, SSR1, TRPC1, PDHX, FZD1, NCOA1, USP13, EFNB2, ACVR1B, COL12A1, DAG1, DUSP8, B4GALT5, KIF3B, ARHGEF6, ROCK2, PGGT1B, PIK3R2, EPHA3, ROCK1, EOMES, BCAT1, EFS, DNAJA2, MAN1A1, PDE4A, LYPLA1, CAP2, SLC35A1, GNA13, TCFL5, RALBP1, PIM2, SHOC2, SIRT1, TBK1, MAPRE2, CORO1C, VGLL4, SERTAD2, ARHGAP11A, TBC1D4, JOSD1, CHSY1, STK38L, WAPAL, ZCCHC14, CAMTA1, KIAA1033, ARHGEF4, NBEA, CDC40, YBX2, RNF138, PHF20, KLF3, POGK, TRPM4, PALMD, PRPF40A, DET1, HMG20A, BRWD1, NDFIP2, SLC39A10, DOLPP1, INTS2, KIAA1468, LRRN1, ZFYVE28, ZDHHC6, SMURF2, BCL11A, BCL11B, SLC25A32, ANP32E, KCTD10, ARL6, LCOR, C1orf198, DIRC2, TRIM41, SLC44A1, SOCS4, PANK1, SP1, PPTC7, SLC39A13, LONRF1, ACOT4, MIER3, YTHDF3, DGKH, SLC9A9, IDH3G, UBR1, IL6ST, SPTBN1, MTDH, ATXN1, OTC, CACNA1D, NR3C2, SMAD5, PTER, MTUS1, GPM6B, GOLGA7, OSBPL8, SP3, VLDLR, TNRC6B, CD47, CSNK1A1, CD68, TMEM70, TTC14, MYO1C, PLAGL1, ZBTB34, RAP1GDS1, PCMTD2, ANK3, ARHGAP6, BACH1, PRDM1, LDB2, EMP1, HIF1A, ALCAM, ATP1B1, FRK, GNAQ, KPNA3, MAN2A1, PDGFA, PLAG1, PTPN1, RAB5B, SKI, RPS6KB1, ZKSCAN1, CNTNAP1, GAS7, KLF4, COX5A, DUSP5, ESRRA, TGFBR1, AKT3, PTK2, TOPORS, MAT2A, SEMA3A, TXNIP, API5, AHCYL1, CPLX1, QKI, TMED10, UTRN, ANGPTL2, BACE1, TLK1, TNPO2, BZW2, ORMDL2, CDR2L, RGL1, RCOR1, PSD3, GULP1, TMEM9, Mar-05, ZNF654, ST7, NUDT4, SMEK2, PELI1, CRAMP1L, PTBP2, PELI2, SNX16, NUCKS1, ELOVL6, SPSB1, SETD7, RAB1B, SEH1L, ZNRF3, SLITRK6, ASPH, PURB, CTTNBP2, MBD6, NAGS, AEBP2, FBXL16, FAM84B, CHMP4B, BCL9L, KCTD1, TSEN54, CAPN3
hsa-miR-628-	BTK, PTEN, ATXN1, ATRX, IGF1R, PPP3CA, ZCCHC11, Sep-07, DCTD, BTBD11, CUL4B, MEX3A, PCDH19, CPD, MAPK6, TCF7, PIP5K1B, BPTF, DLG5, CCDC6, RBL2, YAF2, PAIP1, ZBTB6, TLK1, ZNF516, KIAA0922, KBTBD2, LSM14A, LUC7L2, NLK, CAMK2N1, EPB41L4B, SPIRE1, NUFIP2, PPP3CB, PTBP2, FAM60A, FRY, LSM12, MIER3, PAN3, CREM, AMPD3, SAMD13, TNRC6B, PLEKHG1, TFAP2A, PCDH17, CYLD, TRIM23, BPGM, CSNK1D, GCLM, HLF, ISL1, KPNA1, PPP2R1B, RPS6KB1, KIF3B, MAP4K4, GPM6A, SEC23A, ADAMTS1, CBX3, FBXL3, TRPS1, VAMP2, ZBTB11, ZBED4, SIN3A, TRIM33, UBE2J1, LIMA1, RSBN1, EML4, SLC39A10, DCUN1D1, SLAIN2, RNF38, SMC6, ITCH, CHKB, BTF3L4, MAMDC2, JAZF1, SUPT3H

hsa-miR-15b	ADRB2, CLCN5, FGFR2, INSR, MYO5A, PTCH1, PEX5, PAFAH1B1, APP, RARB, SMAD5, RFWD2, PLEKHA1, ACTR2, EDA, CTNNBIP1, TSC22D3, TMCC1, ZHX1, SERBP1, BCL7A, MINK1, AMMECR1, ARHGAP5, USP14, PCDH17, SIDT2, PTPRD, SLC9A6, SLC12A2, NCOR2, SHANK3, MYO1C, SCN3A, SMARCD2, TMEM100, PHACTR2, PID1, UBE4B, CALU, CAMK2G, CCNE1, CCNT2, CHEK1, CNN1, ESRRG, ACP2, CLCN3, E2F3, EIF5A, EZH1, GABPA, IHH, KDR, LAMC1, MAP1A, PIM1, PLAG1, PLRG1, PSMD7, PTPN3, PTPN4, CX3CL1, SSR1, STC1, TRPC1, CUL2, CBX4, PPAP2A, PPAP2B, CDK5R1, FUBP1, WASL, FASN, FKBP1B, FKBP5, COPS2, TRIP10, ARHGDIA, COL12A1, DACH1, DVL1, EGR3, ACSL4, HDGF, KIF5B, RPS6KA3, STX1A, AXIN2, KIF23, TBPL1, SPAG7, ETV1, CRKL, GLUD1, HAS2, PCMT1, HSPG2, IRS1, PDIA6, SMAD7, MAP3K4, MEOX2, CD164, RBM12, SEMA3A, USP15, SPTLC1, POLR3F, DYNLT3, USP3, EXOC5, BTG2, RBBP6, SLC2A3, SOS2, TAF5, WWP1, UBL3, NUP50, PXMP4, ACOT7, SHOC2, AP3M1, CHORDC1, LMOD1, EPB41L1, RYBP, HEY2, MAPRE1, MMD, NRBP1, NRG1, HSPA4L, CAPN6, TRAM1, SESN1, PDCD4, LRIG2, COBLL1, RAB11FIP2, DZIP1, WDR47, WAPAL, KIF1B, ARHGEF9, CAMSAP1, KBTBD2, LRIG1, UBE2J1, RAB4B, ACSL5, RNF138, PHF20, SLC22A17, APLN, SIX4, BAIAP2, KIF21A, CNNM2, TASP1, Mar-05, RNF125, CDCA4, IPO9, ZNF532, RSBN1, EPB41L4B, YOD1, ETNK1, ANKIB1, PLEKHA5, CYP26B1, FEM1C, CCDC47, SLC39A10, SALL4, GALNT1, DMTF1, RAP2C, BACH2, SEC24A, EGLN1, LPPR2, KLC2, BCL11B, CHAC1, ATP13A3, SLC25A22, DHDDS, DICER1, SLC41A2, SLITRK6, KIAA1804, RSPO3, MAP3K9, PURB, ARHGAP18, CDC42, AMOTL1, ZAK, ZNRF2, ANKRD13B, FAM81A, C9orf69, EED, PIK3R1, OGT, ARRDC4, CARM1, GHR, FKBP1A, GRIN1, HTR4, PDE3B, RARG, SIAH1, BRWD1, CSDE1, Sep-02, ENAH, LPHN1, ZC3H12B, FNTA, MYADM, TNRC6B, USP9X, DCUN1D4, CYLD, TTC14, PVRL2, ZNF436, IRF2BP2, FNDC3A, AATK, MYBL1, USP6NL, ZBTB34, PTAR1, HIGD1A, XPO7, PHF21A, ZBTB10, ADSS, ANK2, CDX2, CHD2, CPD, EIF4B, ACVR2A, CCND2, CLCN4, FGF2, GRB2, KPNA1, LRP6, PEX13, MAPK9, MAP2K1, PTPRR, SORT1, SALL1, ATXN2, SRPK1, RPS6KB1, TCF3, TGFBR3, WEE1, YWHAH, TAF15, DYRK2, PPM1D, CNTNAP1, AP1S2, BTAF1, SH2D2A, MAP7, SOCS6, ESRRA, MTMR4, MED1, KIF3B, JARID2, RELN, CBFA2T3, MYB, CCDC6, AKT3, DLL1, ABCF2, TSPAN5, RBM6, PURA, RTN3, SMYD5, YAP1, DYNC1LI2, PBX3, IVNS1ABP, GNAI3, SLC20A2, SYPL1, SPTBN2, TLE4, DNAJB4, WIF1, SUPT16H, WHSC1, BACE1, CD2AP, CLDN12, TLK1, PHLDA3, BAZ2A, SOCS5, AP2A1, CARD10, GCC2, HELZ, FRYL, SATB2, TRIM2, SYNE1, CCDC28A, TMEM87A, OSBPL3, PHF19, G0S2, GOLT1B, DCTN4, CAB39, RAPGEFL1, CRIM1, RAB8B, BFAR, GALNT7, RNF111, OTUB1, LRRFIP2, ZCCHC2, CHD7, CDC37L1, STX17, Sep-11, RCOR3, ARHGAP12, ZNF654, TBC1D19, TMEM55A, ASNSD1, CHPT1, CLDN2, DOLPP1, DCUN1D1, JPH1, GATAD2B, USP31, PPM1A, PPP3CB, FAM60A, PELI2, TGIF2, SAV1, SNX16, RNF38, CCNJL, HMBOX1, TBL1XR1, PHC3, KLHL18, C1orf21, SEH1L, GABARAPL1, SH3BGRL2, PHF20L1, PCGF5, SCOC, GPR124, SYDE1, RHPN2, CCND1, CDKN2B, SLC44A1, ZSWIM3, FAM91A1, SPRED1, PDIK1L, RBM24, N4BP1, AEBP2, ODF2, AQP11, C15orf37, CPEB2, RASSF5, ZNF326, KCTD1, E2F7, QKI, CAPN3
hsa-miR-378	FANCA, NR3C1, PAFAH1B1, IGF1R, ATP2B4, GPM6B, FOXP4, ZHX1, PAPD5, GRSF1, KIF2A, Sep-08, RAP1GDS1, PDE4D, BMP2, TCF12, FZD5, FKBP5, CHST2, CALD1, XPR1, ABCF2, DYNC1LI2, RBM14, EXOC5, EPB41L3, RANBP6, DMXL2, AHCTF1, CAB39, CHD8, TRIB2, HSPA12A, PURB, GPT2, FMNL3, CREM, HIPK3, SLC7A6, IRF2BP2, SHANK3, DYRK1A, CELSR3, CENPB, GRB2, HOXB3, PLAG1, MAPK1, PDIA4, VAT1, ZFPM2, KPNA6, AAK1, CDC40, VANGL2, XPO5, NUFIP2, DHX36, KIAA1522, PAPOLA, DCBLD2, PDIK1L, ANKRD52, PTPLB, HDDC3

So for each prognostic miRNA, a list of putative candidate genes was extracted using five different predicting algorithms (table 10) followed by an independent correlation analysis. We focused on only anti-correlated miRNA, mRNA pairs according to the biology of miRNAs. The independent correlation analysis was performed among the pairs due to lack of correlation between the miRNAs and the predicted targets obtained from different algorithms.

Table 11. Table of top anti-correlated miRNAs and their target genes.

miRNAs	**Gene**	**miRNA:mRNA Pearson correlation**	**miRNAs**	**Gene**	**miRNA:mRNA Pearson correlation**
hsa-miR-378*			hsa-miR-330-5p		
	TMEM171	-0.05		DNAJA3	-0.18
	ODC1	-0.03		C16orf54	-0.08
	ACO2	-0.05		TCF7	-0.11
	PAFAH1B1	-0.04		NR0B2	-0.19
	PLOD1	-0.11		CEP250	-0.04
	TGM2	-0.05		TRPC4AP	-0.06
	CTBP2	-0.18		CDK5RAP1	-0.12
	NR3C1	-0.02		EPS8L3	-0.03
	UTP18	-0.15		ZDHHC9	-0.17
	FANCA	-0.12		FAM107B	-0.08
	TCN2	-0.05		RCSD1	-0.03

	EPB41L3	-0.09		LCOR	0.00
	RPN2	-0.04		SHD	-0.11
	HMGB2	-0.13		GLOD5	-0.15
	PTPLB	-0.04	hsa-miR-323-3p		
	KLC1	-0.02		AGT	-0.03
	PRKCDBP	-0.01		E2F1	-0.03
	NAT10	-0.01		HK2	0.00
	SIGLEC1	-0.13		EXO1	-0.08
	SLC26A6	-0.05		TTK	-0.02
	UHRF1	-0.11		DFFB	-0.08
	ACSS2	-0.15		SPC25	-0.05
hsa-miR-15b*				OIP5	-0.08
	ARL6IP5	-0.16		SELENBP1	-0.14
	PAFAH1B1	-0.08		NUSAP1	-0.04
	PPT1	-0.04		TIMM23	-0.01
	SMARCD2	-0.10		E2F8	-0.02
	TOMM34	-0.04		SETD7	-0.07
	ADAM	-0.04		C9orf41	-0.04

	12				
	CD48	-0.05		CMTM8	-0.18
	CXCL10	-0.10	hsa-miR-125a-3p		
	STC1	-0.07		ETV4	-0.04
	ZW10	-0.05		TSEN54	-0.07
	FKBP1B	-0.09		AHCY	-0.18
	NCOR2	-0.05		CTBP2	-0.13
	SMYD5	-0.24		TUBG1	-0.05
	GART	-0.08		POLE2	-0.05
	CHEK2	-0.16		SLC26A3	-0.05
	OIP5	-0.07		CEP250	-0.05
	TGIF2	-0.09		FAP	-0.01
	DUS1L	-0.05		TMEM97	-0.06
	SRPRB	-0.14		DUSP7	-0.16
	FAM60A	-0.17		CSGALNACT2	-0.09
	SLC15A4	-0.19		ZNF703	-0.20
	C10orf54	-0.03		MRRF	-0.05
	ACSS2	-0.11		ZNF511	-0.18
hsa-miR-628-3p				ZNRF3	-0.31
	DNAJA	-0.01	hsa-miR-		

	3		655		
	PKP4	-0.16		DNAJA3	-0.14
	STK17B	-0.11		SUV39H2	-0.08
	GPR19	-0.10		AFG3L2	-0.14
	ASPM	-0.13		CALU	-0.15
	FAM3D	-0.30		WSB1	-0.02
	SAMD13	-0.24		ETV5	-0.04
	MARVELD3	-0.28		CCNA2	-0.15
hsa-miR-135a*				PTTG1	-0.23
	TSEN54	-0.03		BUB1B	-0.04
	RRM1	-0.03		MMP7	-0.20
	TNC	-0.15		TLE4	-0.08
	CHSY1	-0.01		PLK4	-0.09
	DUSP4	-0.04		CCNE2	-0.10
	IL6ST	-0.03		NR3C2	-0.14
	SRPX	-0.11		TNFAIP6	0.00
	TCP11L1	-0.04		CHRNA5	-0.01
	CD36	-0.01		RNF11	-0.16
	DUT	-0.01		SPON1	-0.02
	FAP	-0.01		AHCYL2	-0.16
	COL5A	-0.03		CCPG1	-0.14

	2				
	FAM107B	-0.02		SELENBP1	-0.10
hsa-miR-183*				SLC38A2	-0.14
	TSEN54	-0.01		KIF20A	-0.22
	ACO2	0.00		HSPA14	-0.11
	EBNA1BP2	-0.08		E2F8	-0.11
	NR3C1	-0.02		FAM60A	0.00
	PKP4	-0.07		A1CF	-0.16
	KIAA0101	-0.01		CENPK	-0.18
	TARBP1	-0.08		SH3RF1	-0.19
	PCK2	-0.17		UHRF1	-0.18
	PHYH	-0.08		ANKRD44	-0.05
	DUSP4	-0.12	hsa-miR-92a-1*		
	CXCL10	-0.10		TNC	-0.10
	TLE4	-0.15		RNASE1	-0.09
	LSM6	-0.10		CLDN7	-0.04
	EPB41L3	-0.06		KNTC1	-0.01
	PBXIP1	-0.03		KLC1	-0.12
	HSD17	-0.10		DUS1L	-0.12

	B11				
	FAM107B	-0.13		EXOSC5	0.00
	CCNDBP1	-0.06		TMEM52	-0.04
	GPR34	-0.02			

In order to established relation between miRNAs and their respective mRNA gene targets, we carried out target prediction analysis followed by Pearson correlation analysis. The main purpose of this proposed model is the biology of miRNA and the methods in prediction of gene targets. It has been widely known that the binding of miRNA during transcriptional activity degrade predicted targets, so only the anti-correlated predicted targets can prove to be the real ones. Second, the predicted targets are usually unreliable and cannot be implied in biological observation. Thirdly, only miRNAs showing the anti-correlation between its targets can and are more likely to play a role in functional activities.

We focused on only anti-correlated miRNA, mRNA pairs according to the biology of miRNAs. The independent correlation analysis was performed among the pairs due to lack of correlation between the miRNAs and the predicted targets obtained from the five algorithms. For miRNA:mRNA pairs anti-correlation analysis, we isolated each miRNAs targets predicted by the five prediction algorithms and then extracted expression of the matching target genes from the mRNA expression sets.

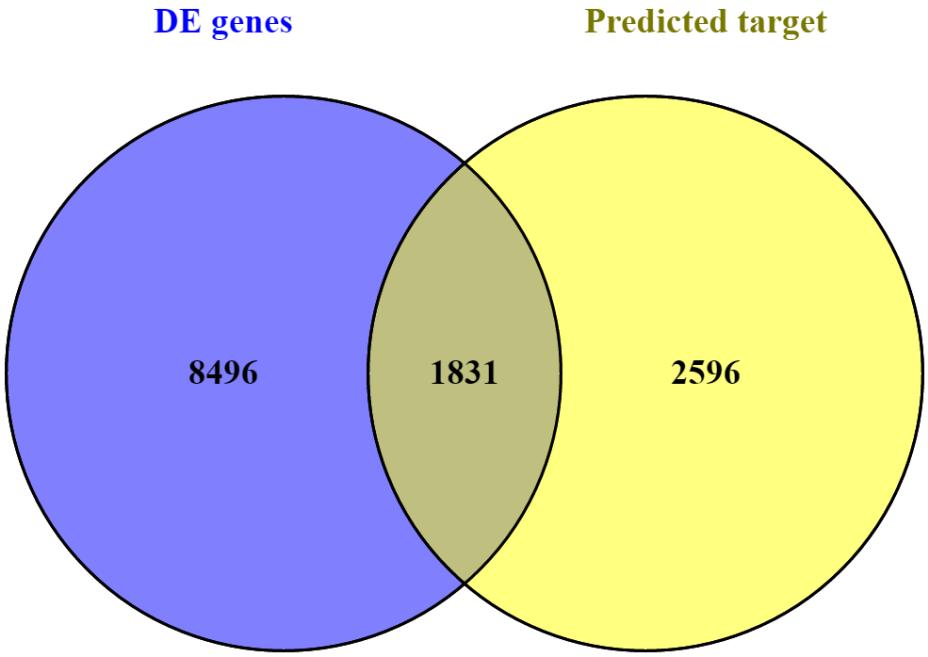


Figure 8. Venn diagram showing common genes among the predicted targets and mRNA expression set.

In total, we observed over-lapping of 1831 (figure 9) genes among the predicted targets and the mRNAs genes differentially expressed among the colon subtypes. The top-anti correlated genes are highlighted in the table (table 11).

4.2.8 Subtyping colon cancer and signatures in mRNA identified by gene expression data

We performed integrated analysis in order to identify miRNAs whose expression is correlated with inverse expression of mRNA targets in primary colon expression set. Therefore, we performed mRNA expression profiling across 585 primary colon samples and identified molecular signatures. In the first step, we performed unsupervised K-mean consensus clustering to uncover potential subtypes of colon tumour on the basis of the similarities of their gene expression

values of 10794 unique genes. We run K=2 to 6 in core K-mean clustering, two molecular subtypes could be identified when K=2 and the cluster consensus was 0.81 and 0.87 for each subtype with 150 and 435 samples. When K=5 (figure 10), the unsupervised clustering reached the highest consensus 0.88 and 0.99. Therefore, we named these subtypes as C1 with 146 samples, C2 with 60 samples, C3 with 232 samples, C4 with 120 samples and C5 with 27 samples.

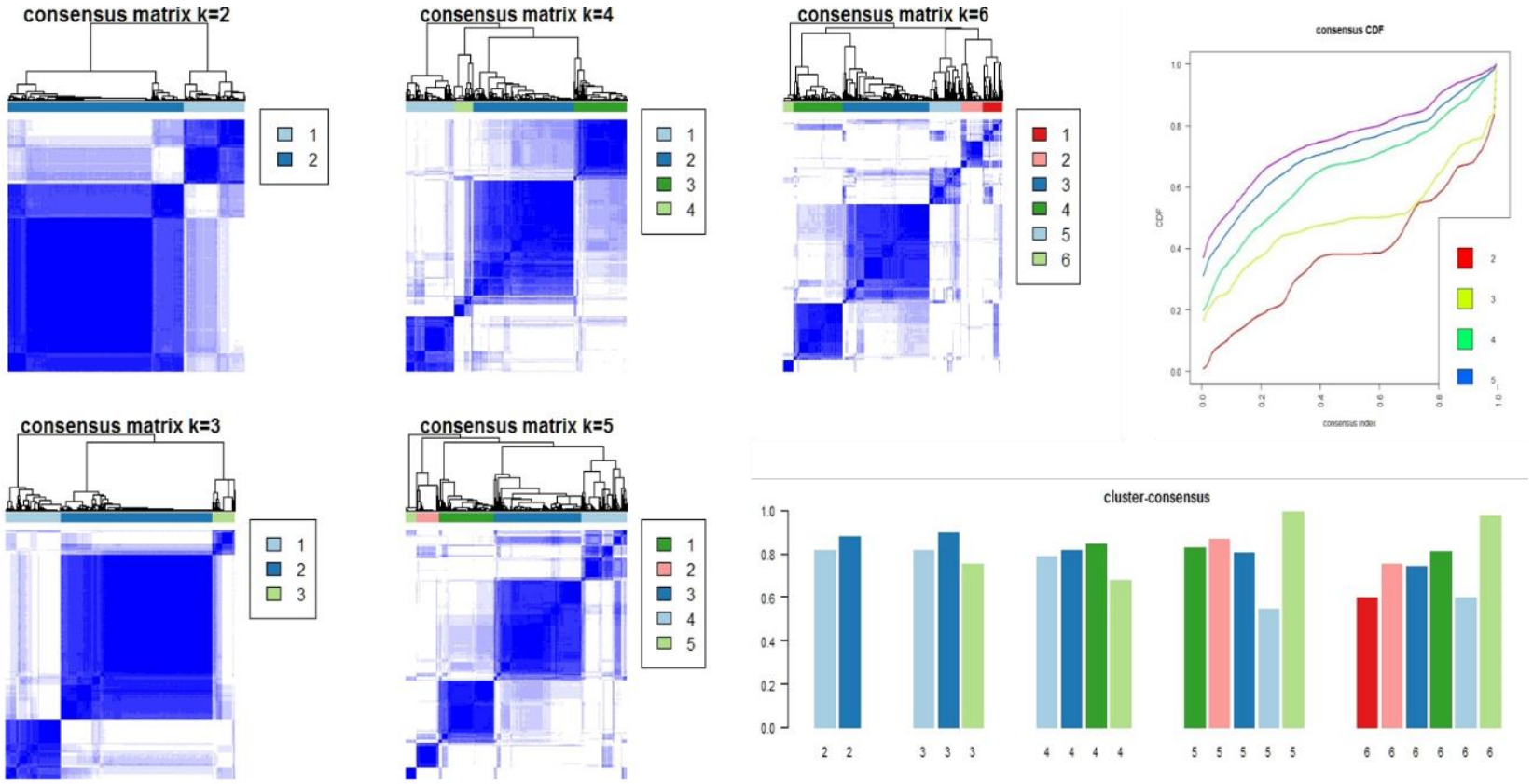


Figure 9. Results from unsupervised K-mean consensus clustering showing running value of K= 2 to 6.

We further carried out differential expression among the five predicted subtypes in order to identify the most discriminating genes among them. We performed ANOVA analysis among the predicted subtypes and selected 716 significant (adj.p.val < 0.001) genes for the two-dimensional average linkage hierarchical clustering. The clustering analysis divided differentially expressed genes into four large sub-groups with different expression subsets represented in the heatmap (figure 11). We further performed gene ontology enrichment analysis for each identified subgroup in order to explore potential cellular processes, molecular functions and biological pathways. The DAVID analysis (table 12) showed that the Cluster I consist of 117 genes was significantly (p-value <0.01) enriched with lipid biosynthesis process, sodium channel regulatory activity, positive regulation of metabolic process, digestion, cellular respiration and inorganic anion transport GO terms. Hierarchical clustering analysis shows that majority of genes from this cluster were up-regulated in C2 and C4 subtypes of colon cancers. We also observed PPAR signalling and mitochondrial carnitine palmitoyltransferase (CPT) system pathways contributed by PPARA, HMGCS2, FABP1, PCK2, CPT1A and ACSS2 genes. Cluster II: is the largest clusters identified by hierarchical clustering consist of 348 genes and were up-regulated in a C5 subtype of colon cancers. We observed over-representation of cellular processes such as cell cycle, metabolic processes, cell division, DNA and RNA replication, cell cycle check points, chromosome organization, DNA repair, ATP binding, DNA and RNA processing, and signal transduction. The cluster genes also showed enrichment of cell cycle, mitotic, cell cycle, DNA replication, pyrimidine replication, DNA repair, CDC20 mediated degradation of Nek2A, the role of BRCA1, BRCA2 and ATR in cancer susceptibility and p53 signalling pathways.

Similarly, Cluster III genes were mostly up-regulated in C1 and C3 subtype of colon cancer. There were 121 genes enriched with intracellular cellular signalling, cellular response, and ion homeostasis GO terms. We also identified Fc Gamma R-mediated phagocytosis pathway with in this cluster. Cluster IV: is consist of 134 highly dysregulated in C1 and C5 subtype of colon cancer. Enrichment analysis shows over-representation cellular processes such as extracellular matrix, biological adhesion, metabolic process, inflammatory response, cell proliferation and cell differentiation, and endopeptidase activity. We also observed ECM-receptor interaction signalling and focal adhesion pathways within this cluster.

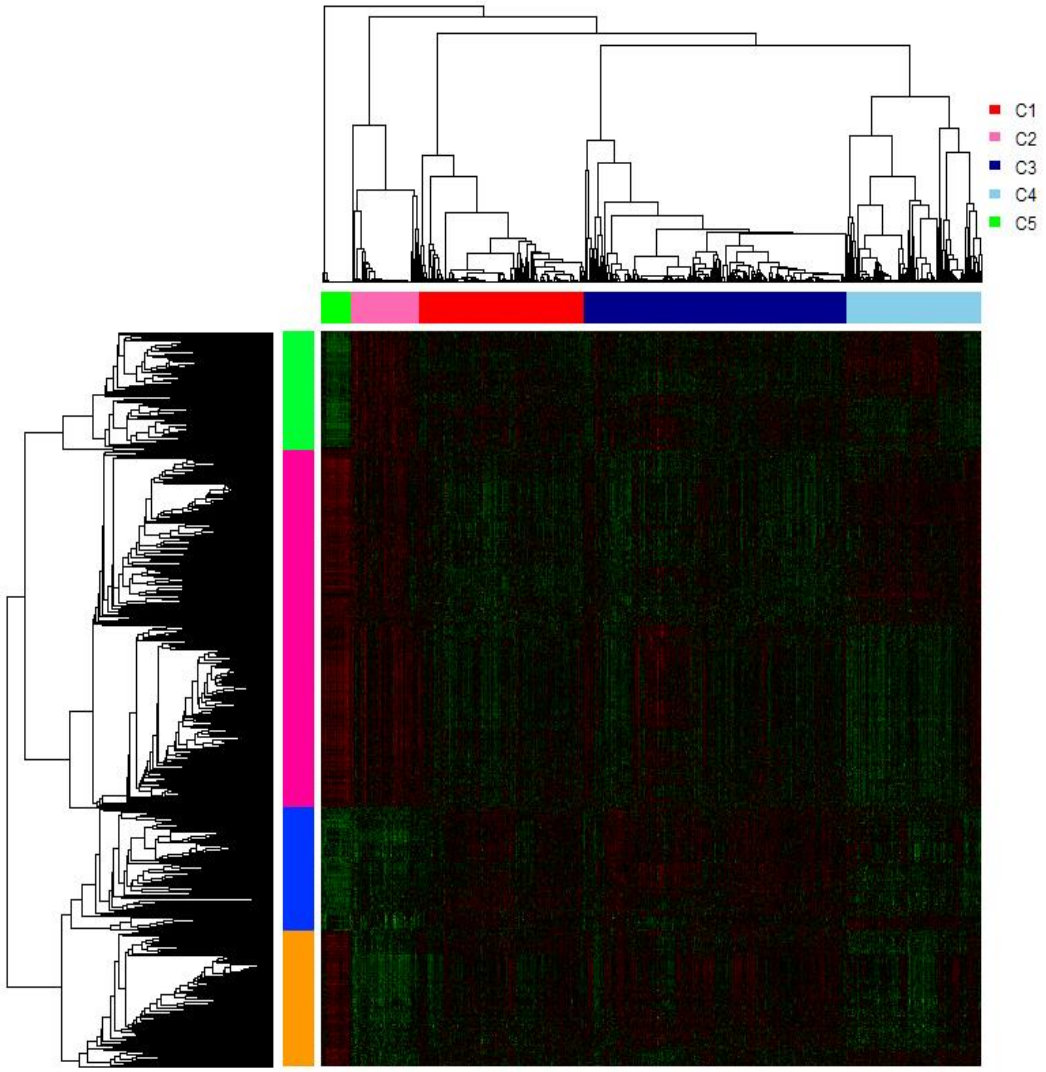


Figure 10. A heatmap showing two-dimensional average hierarchical clustering of five predicted colon subtypes.

We further carried out differential expression among the five predicted subtypes in order to identify the most discriminating genes among them. We performed ANOVA analysis among the predicted subtypes and selected 716 significant (adj.p.val < 0.001) genes for the two-dimensional average linkage hierarchical clustering. The clustering analysis divided differentially expressed genes into four large sub-groups with different expression subsets represented in the heatmap (figure 11). We further performed gene ontology enrichment analysis for each

identified subgroup in order to explore potential cellular processes, molecular functions and biological pathways.

Table 12. Function analysis of four subgroups identified from unsupervised clustering analysis.

Heatmap Clusters	Go terms/pathways
Cluster 1	Lipid biosynthetic process Steroid binding Sodium channel regulator activity Secondary active sulfate transmembrane transporter activity Mitochondrion PPAR signaling pathway Mitochondrial membrane Sulfate transmembrane transporter activity Sulfate transport Mitochondrial envelope Excretion Positive regulation of fatty acid metabolic process Mitochondrial part Mitochondrial inner membrane Organelle envelope Envelope Steroid biosynthetic process Organelle membrane Organelle inner membrane Digestion Inorganic anion transport Aerobic respiration Cellular respiration Mitochondrial carnitine palmitoyltransferase (cpt) system pathway Inorganic anion transmembrane transporter activity
Cluster 2	Mitotic cell cycle Cell cycle Cell cycle phase Cell cycle process M phase Mitosis Nuclear division M phase of mitotic cell cycle Organelle fission Cell cycle, mitotic pathway Intracellular organelle lumen Membrane-enclosed lumen Organelle lumen Nuclear lumen Non-membrane-bounded organelle Intracellular non-membrane-bounded organelle DNA metabolic process Condensed chromosome Cell division Chromosome Spindle Chromosomal part Chromosome, centromeric region DNA replication Condensed chromosome, centromeric region Cell cycle pathway

	Nucleolus Regulation of cell cycle Condensed chromosome kinetochore Response to DNA damage stimulus Nucleoplasm Spindle pole Microtubule cytoskeleton Chromosome segregation Kinetochore Cell cycle checkpoint Microtubule cytoskeleton organization Microtubule-based process DNA repair Mitotic sister chromatid segregation Sister chromatid segregation Chromosome organization Spindle microtubule Regulation of mitotic cell cycle Regulation of cell cycle process Cellular response to stress Mitotic cell cycle checkpoint DNA-dependent DNA replication Spindle organization Cell cycle checkpoints pathway Nuclear chromosome ATP binding Nucleoside binding Adenyl nucleotide binding Adenyl ribonucleotide binding Cytoskeletal part Purine nucleoside binding Interphase of mitotic cell cycle Interphase Macromolecular complex subunit organization DNA replication pathway Pyrimidine metabolism pathway Ribosome biogenesis Purine nucleotide binding DNA replication pathway Ribonucleotide binding Purine ribonucleotide binding Microtubule organizing center Ribonucleoprotein complex biogenesis NCRNA processing DNA integrity checkpoint DNA binding Nucleotide binding DNA packaging DNA strand elongation during DNA replication Microtubule RNA processing Meiotic cell cycle Macromolecular complex assembly Centrosome DNA strand elongation Condensed nuclear chromosome Regulation of organelle organization DNA damage checkpoint Regulation of ubiquitin-protein ligase activity during mitotic cell cycle Ncrna metabolic process

	Protein targeting to mitochondrion Protein localization in mitochondrion Meiosis M phase of meiotic cell cycle tRNA processing Regulation of protein ubiquitination Regulation of nuclear division Regulation of mitosis Regulation of ubiquitin-protein ligase activity P53 signaling pathway Regulation of ligase activity Role of brca1, brca2 and atr in cancer susceptibility pathway Telomere maintenance pathway Cytoskeleton Lagging strand elongation Mitochondrion organization Chromosome condensation Cytoskeleton organization Phosphoinositide-mediated signaling Ribonucleoprotein complex Anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process Negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle Negative regulation of ligase activity Negative regulation of ubiquitin-protein ligase activity Oocyte meiosis pathway Mitotic chromosome condensation Spindle pole body Replication fork DNA repair pathway Chromatin Negative regulation of protein ubiquitination Establishment of chromosome localization Mitotic spindle organization Chromosome localization Protein complex assembly Protein complex biogenesis Microtubule organizing center part Kinetochore microtubule Cell proliferation DNA replication initiation Regulation of cyclin-dependent protein kinase activity Nucleotide-excision repair DNA damage response, signal transduction Apc-cdc20 mediated degradation of nek2a pathway Flap endonuclease activity Condensin complex Positive regulation of protein ubiquitination Nuclear chromosome part Mitochondrion Nucleotidyltransferase activity Nuclease activity Mitochondrial matrix Mitochondrial lumen Double-strand break repair RNA polymerase activity DNA-directed RNA polymerase activity TRNA metabolic process Protein serine/threonine kinase activity Nucleoplasm part

	Cellular macromolecular complex subunit organization Negative regulation of protein modification process Protein targeting G2 phase of mitotic cell cycle G2 phase Organelle localization rRNA processing Regulation of microtubule cytoskeleton organization Microtubule binding Regulation of mitotic metaphase/anaphase transition Positive regulation of protein modification process Positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle RRNA metabolic process Covalent chromatin modification Mitochondrial transport Establishment of organelle localization Biopolymer methylation Establishment of mitotic spindle localization Chromatin assembly or disassembly Positive regulation of ubiquitin-protein ligase activity Purine metabolism pathway Helicase activity Microtubule motor activity Brca1-dependent ub-ligase activity pathway Phospho-apc/c mediated degradation of cyclin a pathway Chromatin organization Protein import Positive regulation of cellular protein metabolic process Positive regulation of ligase activity Mitotic metaphase plate congression Proteasomal ubiquitin-dependent protein catabolic process Proteasomal protein catabolic process Structure-specific DNA binding RNA modification Regulation of microtubule-based process Methylation Base-excision repair Deoxyribonucleotide metabolic process Chromatin binding Protein kinase activity Spindle localization Establishment of spindle localization Positive regulation of protein metabolic process Nuclear matrix Centriole DNA-dependent atpase activity Cellular component disassembly Cajal body Mitotic cell cycle spindle assembly checkpoint Metaphase plate congression Negative regulation of mitotic metaphase/anaphase transition Base excision repair pathway One-carbon metabolic process Negative regulation of cellular protein metabolic process Positive regulation of organelle organization RNA binding Deoxyribonuclease activity Cellular protein localization Nuclear periphery Regulation of exit from mitosis

	Negative regulation of nuclear division Negative regulation of mitosis Spindle checkpoint Pronucleus
Cluster 3	Di-, tri-valent inorganic cation homeostasis Cation homeostasis Cellular di-, tri-valent inorganic cation homeostasis Cellular cation homeostasis Intracellular signaling cascade Fc gamma r-mediated phagocytosis pathway Ion homeostasis Response to organic substance Cellular calcium ion homeostasis Calcium ion homeostasis Cellular metal ion homeostasis Cellular ion homeostasis Cellular chemical homeostasis Metal ion homeostasis Chemical homeostasis Response to wounding
Cluster 4	Proteinaceous extracellular matrix Extracellular matrix Extracellular region part Extracellular region Extracellular matrix part Response to wounding Collagen Cell adhesion Biological adhesion Extracellular space Calcium ion binding Collagen metabolic process Multicellular organismal macromolecule metabolic process Basement membrane Inflammatory response Multicellular organismal metabolic process Ecm-receptor interaction pathway Smad binding Protein dimerization activity Metalloendopeptidase activity Skeletal system development Blood vessel development Collagen catabolic process Vasculature development Regulation of cell proliferation Muscle cell differentiation Defense response Multicellular organismal catabolic process Collagen fibril organization Protein heterodimerization activity Focal adhesion pathway Endopeptidase activity

The DAVID analysis (table 12) showed that the Cluster I consist of 117 genes was significantly (p-value <0.01) enriched with lipid biosynthesis process, sodium channel regulatory activity, positive regulation of metabolic process, digestion,

cellular respiration and inorganic anion transport GO terms. Hierarchical clustering analysis shows that majority of genes from this cluster were up-regulated in C2 and C4 subtypes of colon cancers. We also observed PPAR signalling and mitochondrial carnitine palmitoyltransferase (cpt) system pathways contributed by PPARA, HMGCS2, FABP1, PCK2, CPT1A and ACSS2 genes. Cluster II: is the largest clusters identified by hierarchical clustering consist of 348 genes and were up-regulated in a C5 subtype of colon cancers. We observed over-representation of cellular processes such as cell cycle, metabolic processes, cell division, DNA and RNA replication, cell cycle check points, chromosome organization, DNA repair, ATP binding, DNA and RNA processing, and signal transduction. The cluster genes also showed enrichment of cell cycle, mitotic, cell cycle, DNA replication, pyrimidine replication, DNA repair, CDC20 mediated degradation of Nek2A, the role of BRCA1, BRCA2 and ATR in cancer susceptibility and p53 signalling pathways.

4.2.9 Prognostic miRNAs and their impact on signatures of colon subtypes

To gain further understanding into the biological impact of the identified prognostic miRNAs, we investigated their relationships with their putative targets and interaction pathways with in the mRNA expression data. We calculated the correlation between miRNA expression and their putative targets across all the tumour samples and identified top miRNA:mRNA pairs. In our results, we observed a clear separation between “protective” and “risk-associated” miRNAs and their associated target genes and their functional pathways.

We identified five prognostic miRNAs (hsa-miR-628-3p, hsa-miR-135a*, hsa-miR-183*, hsa-miR-330-5p, and hsa-miR-655) linked to “risk-associated” or

worse outcome target genes from C4 and C5 colon subtype. Enrichment analysis of genes in C4 and C5 subtypes have shown that unfavourable prognostic miRNAs are broadly related to cell cycle processes important for apoptosis and anti-proliferative activities such as DNA replication, DNA repair, DNA binding, ATP binding and pyrimidine binding. Other important pathways related to pyrimidine metabolism, oocyte meiosis, focal adhesion, ECM-receptor interaction, p53 signalling and telomere maintenance, a significant sub-pathway associated with colon cancer risk (Slattery et al., 2015b).

Similarly, we observed five miRNAs (hsa-miR-378*, hsa-miR-15b*, hsa-miR-323-3p, hsa-miR-125a-3p, and hsa-miR-92a-1*) associated with "protective" outcome have shown correlation with the up-regulatory genes of C5 and C2 subtype of colon cancer. The biological theme positively correlated with favourable prognostics miRNAs were sodium channel regulatory activity, transmembrane transporter activity, cellular respiration, PPAR signalling pathway, cell division, DNA replication, DNA repair and metabolism.

4.2.10 Subtype-specific miRNAs and their impact on transcriptional phenotypes

We investigated the expression of miRNAs among the different types (primary tumour and metastatic) of colon cancer in order to understand the phenomenon behind the establishment of a tumour transcriptional phenotypes. As discussed above, among the 17 dysregulated miRNAs between primary colon tumour and metastatic, 11 miRNAs have shown highest expression among the metastatic tissues. Notably, four prognostic miRNAs (has- miR-15b*, hsa-miR-378*, hsa-miR-323-3p, hsa-miR-125a-3p) were up-regulated among a metastatic tumour and observed with favourable prognosis. All four miRNAs targets genes from cluster I

and II, modulating important processes of transmembrane activity, PPAR signalling, DNA replication and DNA repair.

Likewise, the two prognostic miRNAs (hsa-miR-135a*, hsa-miR-655) were highly up-regulated in primary colon tumours associated with worse prognosis, targets genes from cluster II and IV controlling important processes of cell cycle such as DNA replication, binding and repair, and p53 signalling pathway.

4.3 Discussion

Prognostic stratification of the colon in clinical practice on the basis of gene expression has been proven unreliable to date largely due to the heterogeneity of colon cancer. Another challenging factor among colon cancers is to determine phenomenon behind molecular alterations of primary colon cells towards metastatic capability. In this study, we have performed robust classification of mRNA expression of a colon tumour on the basis of the transcriptome to improve existing stratification of colon cancers and identify subtype-specific signatures targeted by a new class of regulators known as miRNAs. We further investigated miRNAs association with the possible outcome, targeted molecular pathways and novel biological markers for chemotherapy.

We first performed differential expression analysis on the entire cohort of miRNA colon tumour on the bases of histopathological groups in order to stratify the colon cancers. In results, we identify 70 highly dysregulated genes that can differentiate between the primary colon and metastatic tissue classes, colon cancer on the bases of tumour grades, colon cancer on the basis of stage and colon cancers with adjuvant chemotherapy. Subsequently, we identified 10-miRNA signature that can distinguish between the “protective” and “risk-associated” prognostic miRNAs, will required further validation on independent data set. For the primary tumour mRNA, we first sub-classify 585 colon tumours into five subtypes based on expression similarities and then clustered the most discriminatory genes into four signatures. Through, target prediction analysis, we identified that the prognostic miRNAs target genes of these four signatures and control their involvement in molecular processes of the cell cycle and signalling pathways.

4.3.1 The miRNAs expression in primary and metastatic colon cancer

11 miRNAs upregulated in metastatic versus primary tissues strictly represent metastatic pool of colon cancers for which the publicly available software mirPath (Vlachos et al., 2012) identified number of known biomarkers genes and modulate important pathways such as Neurotrophin signaling, MAPK signaling, TGF-beta signaling, Axon guidance, PI3K-Akt signaling, mTOR signaling, ErbB signaling, Chronic myeloid leukemia, Insulin signaling, p53 signaling and Colorectal cancer pathways. Therefore, the observed association of metastatic miRNAs with a large number of signalling pathways confirms the presence of metastatic lymph node RNAs and provide strong evidence of their potential role in the development of metastases. Among the dysregulated miRNAs, hsa-miR-15b is an interesting one and highly upregulated in metastatic compared to primary tissues. We have observed an association of hsa-miR-15b with high-risk survival in this study, modulate p53 tumour suppressor signalling pathway important for high metastatic potential and metastatic relapse.

Similarly, 6 other miRNAs upregulated in primary tissues target functional pathways such as PI3K-Akt signaling, Focal adhesion, Ubiquitin mediated proteolysis, Regulation of actin cytoskeleton, Prostate cancer, mTOR signaling, Pathways in cancer, Melanoma, TGF-beta signaling, and ErbB signaling related pathways and thus expected to be involved in promoting proliferation and inhibiting apoptosis, and also play an important role in tumour microenvironment (Qiu et al., 2016, Aminuddin and Ng, 2016). Most of the miRNAs dysregulated have shown log fold change > 0.5 and some of them have been previously observed in relation to colon cancers such as hsa-miR-135a*, hsa-miR-26a, and

hsa-miR-335. However, mostly miRNAs associated with survival and metastases were not found in association with colon cancer before.

4.3.2 miRNAs as prognostic markers in colon cancer

Using univariate and multivariate Cox analysis we examined the correlation of differentially expressed miRNAs in colon tumour samples with OS and DFS. As a result, we observed 10 miRNAs prognostic signature for OS and DFS with "protective" and risky outcomes. The correlation analysis of prognostic miRNAs with targeted pathways provides indirect (independent) evidence of their association with the biology of a colon tumour. One of the themes related to risk-associated miRNAs were the appearance of functional pathways such as TFG-beta, T cell, Wnt, mTOR and ErbB signalling pathways vital for cell cycle progression, proliferation, apoptosis, differentiation and migration (Oh et al., 2016, Aminuddin and Ng, 2016). The other emerging theme related to protective roles of miRNAs were the presence of cell cycle and related processes important for inducing apoptosis, cell cycle arrest and inhibition of cell migration (Wu et al., 2016, Hu et al., 2016).

We then focused on individual miRNAs for their impact on colon tumour microenvironment. Risk-associated miRNAs such as hsa-miR-628-3p, previously reported in association with a diagnostic marker for low-stage pancreatic cancer (Li et al., 2013) is highly upregulated among the stages comparison particularly in early stages (Stage I and II) of a colon tumour in our study. Integrated analysis demonstrated that the top anti-correlated targets genes for hsa-miR-628-3p were DNAJA3, PKP4, STK17B, GPR19, ASPM, FAM3D, SAMD13, and MARVELD3, belongs to Cluster I and II functional subgroups and specifically upregulated in C2, C4 and C5 subtypes of colon cancer. These genes are

potentially involved in important regulatory processes such as intracellular organelle lumen, membrane-enclosed lumen, organelle lumen and pathways of p53 signalling, PPAR signalling pathway, cell cycle and DNA repair. hsa-miR-323-3p previously reported in association with metastases in pancreatic (Wang et al., 2016) and cervical squamous cell carcinoma (Ding et al., 2014) has also shown over-expression metastatic tissue. From the integrated analysis, we also predicted gene targets AGT, E2F1, HK2, EXO1, TTK, DFFB, SPC25, OIP5, SELENBP1, NUSAP1, TIMM23, E2F8, SETD7, C9orf41 and CMTM8, involved in important processes of apoptosis, cell cycle, metabolism and mismatch repair. The analysis shows downregulation of all these predicted target genes and upregulation of hsa-miR-323-3p in metastatic tissue explains its "protective" role; most of these target genes belong to a C5 subtype of colon cancer. Another microRNA, hsa-miR-125a-3p predicting OS is upregulated in metastatic tissue defined as "protective". Till date, nothing has been reported of hsa-miR-125a-3p in association with colon cancer but has been involved in multiple abnormalities (Huat et al., 2015, Bi et al., 2015, Tang et al., 2015). However, hsa-miR-125a-3p have been involved in its protective functions by increased apoptosis (Ninio-Many et al., 2014) and a wide range of biological processes including regulation of Wnt/beta-catenin signalling pathway (Choi et al., 2011). The upregulation of hsa-miR-655 in metastatic tissues linked to "risk associated" overall survival. hsa-miR-655 has been reportedly involved in a number of tumorigenesis and drug resistance. hsa-miR-655 regulate TGF-β-induced epithelial-mesenchymal transition, a key element of cell invasion, migration, metastases and drug resistance. Anti-correlation analysis shows that patients with upregulation of hsa-miR-655 and downregulation of its targets mRNA genes carry the risk of

metastases. The potential target genes of hsa-miR-655 belong to subtype C2, C4 and C5 of colon cancer.

Prognostic marker for DFS, hsa-miR-92a-1* dysregulated among the histological grades, particularly upregulated among poor grades defined as "protective" in action. hsa-miR-92a-1* has been previously reported in connection with colorectal cancer by consistent amplification of MIR17HG, CMYC, and ABCC4 genes (Molinari et al., 2016). The anti-correlation analysis shows that hsa-miR-92a-1* targets TNC, RNASE1, CLDN7, KNTC1, KLC1, DUS1L, EXOSC5, and TMEM52 of C5 subtype of colon cancer, involved in processes like cell adhesion, RNA degradation and ECM-receptor interaction.

4.3.3 How miRNAs expressed in different pathological groups?

We have observed 70 miRNAs differentially expressed among the histological groups. We observed the highest number of dysregulated genes when we compared expression data among the three different grades. As anticipated, we observed a theme that those miRNAs which have shown upregulation among the metastatic tissues tend to over-expressed among the poorly differentiated grades and higher stages (III and IV). Such as hsa-miR-378* miRNA have shown higher expression among the metastatic tissue compared to primary colon has also highest expression level among the poorly differentiated grades and Stage III & IV. Likewise, other miRNAs such as hsa-miR-424* has also shown similar trends. The identified findings from this study are in agreement with independent studies (Schneider and Langner, 2014, Derwinger and Gustavsson, 2008) therefore shows the validity of confirmed the validity of histological based feature selection. These finding also reinforced the idea of RNA based classification at the miRNA level.

We have also found some of the miRNAs expression level was very specific to one histological group, indicate histological based targeting of mRNA and oncogenic pathways. We have found 7 miRNAs specifically belongs to primary colon versus metastatic group and most of them have shown downregulation in metastatic tissues. Further investigation on individual miRNA may highlight factors behind metastases. Similarly, we identify 12 miRNAs specific to Stages comparison only and most of these miRNAs have shown over-expression among the early Stages (I and II) of colon tumours. Therefore, stage level based miRNA-mRNA associations can be major contributors of different transcription phenotypes.

In conclusion, we have performed miRNA, mRNA integration with pathological and clinical information of well-characterised cohorts of colon cancers. This study present advance histological based classification of miRNA expression data and their role in regulating subtype-specific transcriptional signatures. Furthermore, we have classified primary colon into five major subtypes on the bases of their RNA levels and also divided five subtypes and four functional groups on the basis of their involvement function pathways. We have also provided a dissection of aberrations at the miRNA level, their impact on targets and perturbations among functional pathways.

4.4 Methods

4.4.1 Data collection and preprocessing

Both expression data sets consist of miRNA and mRNA profiles were extracted from GEO (Gene Expression Omnibus). miRNA expression was extracted from FFPE colon tissues for microRNA array analysis using NIH Taqman Human MicroRNA Array v.2 platform can be found under the accession number GSE29622. The miRNA expression set consist of 47 primary and 18 metastatic colon cancers with patient follow-up and extensive histopathological information. Likewise, mRNA expression data set consists of 598 colon cancer samples analysed using Affymetrix U133plus2 chip profiles can be found under the Subseries accession number GSE39582.

We applied functions of Bioconductor GEOquery package (Davis, 2013) for the extraction of raw expression data for both types of RNA sets. Various functions of Robust Multichip Analysis (RMA) methodology were applied for background correction followed by quantile normalisation for the correction of inter-arrays global differences. All the probes with Zero variance were filtered for further analysis. R, Hclust functions were used for the calculation of two-dimensional average-linkage hierarchical clustering and heatmaps were drawn using rows as scale.

4.4.2 Statistical analysis

We applied Cox-regression analysis for the calculation of association of miRNA level expression with disease-free survival (DFS) and overall-survival (OS) followed by adjustment of wald test p-values using multiple-testing using Benjamin-Hochberg method. We consider all those miRNAs were having

threshold value less than 0.05 associated with DFS and OS. Survival data were censored at the date of alive/death from any cause for OS and the evidence of recurrence or no recurrence for DFS. We applied different Cox-regression models for histopathological covariates, grouped as Gender (male or female), Tumour Type (a primary tumour versus metastatic), Grades (continuous), Stage (continuous) Adjuvant Chemotherapy (no or yes). Survival curves were calculated and drawn by Kaplan-Meier analysis using R analysis "survival" tools.

4.4.3 Identification of dysregulated miRNAs linked to specific colon subtypes

We applied ANOVA and on some instance Student's t-test for the identification of differentially expressed miRNAs across the different histological groups such as (primary versus metastatic) (well versus moderate versus poorly grades), among the stages, and (adjuvant chemotherapy yes versus no). All the obtained p-values were adjusted for multiple testing using False Discovery Rate using threshold value (q-value < 0.01) for differentially expressed genes.

4.4.4 Collection of miRNAs targets and independent correlation analysis

In order to perform integration analysis, we collected candidate targets for each miRNAs using six different databases (TargetScan, TarBase, PicTar, mirBase, miRTarget2, miRanda) using R Bioconductor package "RmiR". We performed independent correlation analysis among the miRNAs and their particular targets in order to evaluate the influence of each miRNA. We selected the top anti-correlated miRNA-mRNA pairs for further analysis.

4.4.5 Classification of primary colon into subtypes and identification of functional signatures

We performed unsupervised K-mean consensus clustering using 'ConsensusClustering' package of R Bioconductor for the identification of potential subtypes of a colon tumour. We run values starting from K=2 to K=6 in core K-mean clustering. A highest consensus score was used for the selection of a number of classes followed by ANOVA analysis for the identification of dysregulated genes among the intrinsic subtypes. Again, p-values were adjusted using False Discovery Rate and mRNA genes exhibiting q-value less than 0.001 termed as significant.

4.4.6 Functional analysis

In order to identify perturbed functional pathways of gene signatures, we applied functional classification tools of available databases such as KEGG (Kanehisa et al., 2016), Reactome (Fabregat et al., 2016) and DAVID (Huang et al., 2009a). Threshold p-value less than 0.001 was used for the selection of significant processes, molecular functions and pathways.

CHAPTER 5 CONCLUSIONS AND FUTURE WORK

5.1 **Conclusions**

The main objective of this report is to discuss the latest development involving the application of DNA microarray on colon cancer and to address the key problems in precise and early targets for diagnosis, distinctive molecular classes and clinical outcome. The discovery of DNA microarray has offered tremendous tools such profiling of gene expression which expanded exponentially in the last decade. DNA microarray also showed enormous impact on the identification of gene expression signature vital for early predictions and clinical outcome of multiple cancer types. One of the major challenges in microarray data analysis is a small number of samples size in given study compare to a large number of driving variable of cancer. Small sample size not only made interpretation difficult but model development, considering the heterogeneous nature of cancer. So, it is not surprising that there is an only small convergence between the gene signatures from different investigators discussing colon cancer.

In chapter 2 of the thesis, we highlighted major advancement in the area of colon cancer. We showed that number of studies have been conducted in order to classify colon tumours into subtypes and novel drug targets for the clinical use. We highlighted the earliest method used in the selection of therapeutic targets for colon cancer was the discovery of gene-drug correlation. We have also seen that colon cancer was subjected to classification on the basis of phenotypes based on microsatellite instability (MSI), phenotypes based on the genetic aberrations were presence of genes such as KRAS or BRAF, and phenotypes based on functional pathways were presence of Wnt/ß-catenin, TGF-ß, MAPK, and PI3K signaling.

We also discussed that most of the studies were conducted focusing on individual gene

targets rather than covering all other aspects of heterogeneity. Some of the author in the past have utilised molecular level similarities in the identification of specific pathways affected in diseases, in the identification drug targets and designing of survival outcome classifiers. One of the major contribution offered by a group of Cancer Genome Atlas by performing genome-scale analysis of 276 patients. As result, they observed a number of important genes and critical pathways required for the initiation and progression of colon cancers. Some of the significant findings from this study was the discovery of P53, PI3K, RAS-MAPK, TGF-β, WNT, and DNA mismatch repair pathways. In spite of such progress, there are still some unknown genetic and genomic changes which play a significant role in colon tumorigenesis.

In chapter 3 of the thesis, the identification of molecular markers with prognostic value in colorectal cancer is a challenging task that is needed to define therapeutic guidelines. Despite recent advances in the screening, diagnosis, and treatment of colorectal cancer, an estimated 608, 000 people die every year from this form of cancer, which is 8% of all cancer deaths. We performed two staged integrated bioinformatics analytics on gene expression data sets of three latest developed studies of colon cancer. We identified two groups of integrated signatures from the comparison of normal versus a tumour and tumour versus meets patient's samples. Functional analysis of the diagnostics 267-genes shows over-representation of signaling-related molecules and also significantly involved in cancers related regulatory pathways. The metastatic 124-gene signature shows functionally involved in immune-response, lipid metabolism and PPAR signalling

pathways. Kaplan-Meier estimates of 124-genes using independent data sets shows that higher grade/stage patients have significantly better overall-survival (p=0.001,

HR=2.61(CI 1.43-4.79)) and disease-specific survival rate (p=0.00, HR=2.41(CI 1.28-4.53)) compare to low grade patients. Further biological validation of genes identified in this study may provide vital biomarker targets for colon cancers.

In chapter 4 of the thesis, we have performed miRNA, mRNA integration with pathological and clinical information of well-characterised cohorts of colon cancers. This study present advance histological based classification of miRNA expression data and their role in regulating subtype-specific transcriptional signatures. Furthermore, we have classified primary colon into five major subtypes on the bases of their RNA levels and also divided five subtypes and four functional groups on the basis of their involvement function pathways. We have also provided a dissection of aberrations at the miRNA level, their impact on targets and perturbations among functional pathways.

In summary, we have identified robust and reliable signatures of miRNA and mRNA along with the identification of distinct subtypes of a colon tumour. Another advantage of this study is that we have uncovered genomic features which may have been remained undetected in individual studies and have an important role in tumour progression.

5.2 Future Work

The data integration models have produced promising results for the colon cancer diagnostic and prognostic signature identification. The two independent studies focusing on tissue-based integration method and integration of two different levels

have established workflow models which can be applied on any independent population. Our main focus was to identify diagnostic and prognostic signatures by integrated analysis without losing potential signature genes.

In a complex disease such as colon cancer, there are the only limited amount of microarray data generated from some platform from different studies addressing a similar question. In this situation, we can have considered tissue-based integration method for different generations of the same microarray technology or for the data generated from multiple microarray platforms. The models proposed in this thesis can then be applied to the integrated data sets of the list of common genes to increase statistical power and to derive reliable gene signatures. In this way, we gain statistical power at the price of loss of potential signature genes.

By using the miRNA-mRNA integrative analysis model, we can integrate two different Omics data levels with clinical and pathological data for the accurate prediction of phenotypic outcome. This could lead to improved cancer prognostic signatures which are mixtures of epigenetic factors, gene expression and clinical parameters. Furthermore, other high-throughput data, such as single nucleotide polymorphism, copy number variation, protein expression data, structural data and tissue microarray data, can be effectively combined into the integrated microarray data in similar manners to correlate changes in gene expression profiles with changes in proteomic or phenotypes.

REFERENCES

1. Perez-Villamil, B., et al., *Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behaviour.* BMC Cancer, 2012. **12**(1): p. 260.
2. Greenlee, R.T., et al., *Cancer statistics, 2000.* CA Cancer J Clin, 2000. **50**(1): p. 7-33.
3. Marisa, L., et al., *Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value.* PLoS Med, 2013. **10**(5): p. e1001453.
4. Hutchins, G., et al., *Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer.* J Clin Oncol, 2011. **29**(10): p. 1261-70.
5. Popat, S., R. Hubner, and R.S. Houlston, *Systematic review of microsatellite instability and colorectal cancer prognosis.* J Clin Oncol, 2005. **23**(3): p. 609-18.
6. Zlobec, I., et al., *Clinicopathological and protein characterization of BRAF- and K-RAS-mutated colorectal cancer and implications for prognosis.* Int J Cancer, 2010. **127**.
7. Zhang, B., *Targeting the stroma by T cells to limit tumour growth.* Cancer Res, 2008. **68**(23): p. 9570-3.
8. Yamasaki, M., et al., *The gene expression profile represents the molecular nature of liver metastasis in colorectal cancer.* Int J Oncol, 2007. **30**.
9. Wang, Y., et al., *Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer.* J Clin Oncol, 2004. **22**(9): p. 1564-71.
10. Tran, B., et al., *Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer*, in *Cancer*. 2011.
11. O'Connell, M.J., et al., *Relationship between tumour gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin.* J Clin Oncol, 2010. **28**(25): p. 3937-44.
12. Eschrich, S., et al., *Molecular staging for survival prediction of colorectal cancer patients.* J Clin Oncol, 2005. **23**(15): p. 3526-35.
13. Kang, G.H., *Four molecular subtypes of colorectal cancer and their precursor lesions.* Arch Pathol Lab Med, 2011. **135**(6): p. 698-703.
14. Jass, J.R., *Classification of colorectal cancer based on correlation of clinical, morphological and molecular features.* Histopathology, 2007. **50**(1): p. 113-30.
15. Liotta, L. and E. Petricoin, *Molecular profiling of human cancer.* Nat Rev Genet, 2000. **1**(1): p. 48-56.
16. Brown, P.O. and D. Botstein, *Exploring the new world of the genome with DNA microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 33-7.
17. Lonning, P.E., T. Sorlie, and A.-L. Borresen-Dale, *Genomics in breast cancer[mdash]therapeutic implications.* Nat Clin Prac Oncol, 2005. **2**(1): p. 26-33.

18. Southern, E., K. Mir, and M. Shchepinov, *Molecular interactions on microarrays.* Nat Genet, 1999. **21**(1 Suppl): p. 5-9.

19. Golub, T.R., et al., *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science, 1999. **286**(5439): p. 531-7.

20. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

21. Quackenbush, J., *Computational analysis of microarray data.* Nat Rev Genet, 2001. **2**(6): p. 418-27.

22. Sørlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.

23. Bittner, M., et al., *Molecular classification of cutaneous malignant melanoma by gene expression profiling.* Nature, 2000. **406**(6795): p. 536-40.

24. Welsh, J.B., et al., *Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer.* Proc Natl Acad Sci U S A, 2001. **98**(3): p. 1176-81.

25. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.

26. Bhattacharjee, A., et al., *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proc Natl Acad Sci U S A, 2001. **98**(24): p. 13790-5.

27. Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**(2): p. 203-9.

28. Dyrskjot, L., et al., *Identifying distinct classes of bladder carcinoma using microarrays.* Nat Genet, 2003. **33**(1): p. 90-6.

29. Belbin, T.J., et al., *Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays.* Cancer Res, 2002. **62**(4): p. 1184-90.

30. Ono, K., et al., *Identification by cDNA microarray of genes involved in ovarian carcinogenesis.* Cancer Res, 2000. **60**(18): p. 5007-11.

31. Su, A.I., et al., *Molecular classification of human carcinomas by use of gene expression signatures.* Cancer Res, 2001. **61**(20): p. 7388-93.

32. Ramaswamy, S., et al., *Multiclass cancer diagnosis using tumor gene expression signatures.* Proc Natl Acad Sci U S A, 2001. **98**(26): p. 15149-54.

33. Giordano, T.J., et al., *Organ-specific molecular classification of primary lung, colon, and ovarian adenocarcinomas using gene expression profiles.* Am J Pathol, 2001. **159**(4): p. 1231-8.

34. Shen, R., et al., *Eigengene-based linear discriminant model for tumor classification using gene expression microarray data.* Bioinformatics, 2006. **22**(21): p. 2635-42.

35. Bicciato, S., et al., *Pattern identification and classification in gene expression data using an autoassociative neural network model.* Biotechnol Bioeng, 2003. **81**(5): p. 594-606.

36. Li, L., et al., *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.* Bioinformatics, 2001. **17**(12): p. 1131-42.

37. Liu, J.J., et al., *Multiclass cancer classification and biomarker discovery using GA-based algorithms.* Bioinformatics, 2005. **21**(11): p. 2691-7.

38. Mukherjee, S., et al., *Support vector machine classification of microarray data.* 1999.

39. Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.* Nature, 2000. **403**(6769): p. 503-11.

40. Cole, B.F., et al., *Polychemotherapy for early breast cancer: an overview of the randomised clinical trials with quality-adjusted survival analysis.* Lancet, 2001. **358**(9278): p. 277-86.

41. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.

42. Naderi, A., et al., *A gene-expression signature to predict survival in breast cancer across independent data sets.* Oncogene, 2007. **26**(10): p. 1507-16.

43. Wang, Y., et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.* Lancet, 2005. **365**(9460): p. 671-9.

44. Sotiriou, C., et al., *Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis.* J Natl Cancer Inst, 2006. **98**(4): p. 262-72.

45. Beer, D.G., et al., *Gene-expression profiles predict survival of patients with lung adenocarcinoma.* Nat Med, 2002. **8**(8): p. 816-24.

46. Pomeroy, S.L., et al., *Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, 2002. **415**(6870): p. 436-42.

47. Takahashi, M., et al., *Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification.* Proc Natl Acad Sci U S A, 2001. **98**(17): p. 9754-9.

48. Ein-Dor, L., et al., *Outcome signature genes in breast cancer: is there a unique set?* Bioinformatics, 2005. **21**(2): p. 171-8.

49. Tan, P.K., et al., *Evaluation of gene expression measurements from commercial microarray platforms.* Nucleic Acids Res, 2003. **31**(19): p. 5676-84.

50. Mah, N., et al., *A comparison of oligonucleotide and cDNA-based microarray systems.* Physiol Genomics, 2004. **16**(3): p. 361-70.

51. Kuo, W.P., et al., *Analysis of matched mRNA measurements from two different microarray technologies.* Bioinformatics, 2002. **18**(3): p. 405-12.

52. Nimgaonkar, A., et al., *Reproducibility of gene expression across generations of Affymetrix microarrays.* BMC Bioinformatics, 2003. **4**: p. 27.

53. Wang, H., et al., *A study of inter-lab and inter-platform agreement of DNA microarray data.* BMC Genomics, 2005. **6**: p. 71.

54. Orr, M.S. and U. Scherf, *Large-scale gene expression analysis in molecular target discovery.* Leukemia, 2002. **16**(4): p. 473-7.

55. Fearon, E.R., *Molecular genetics of colorectal cancer.* Annu Rev Pathol, 2011. **6**: p. 479-507.

56. Sanchez, J.A., et al., *Genetic and epigenetic classifications define clinical phenotypes and determine patient outcomes in colorectal cancer.* Br J Surg, 2009. **96**(10): p. 1196-204.

57. Iacopetta, B., F. Grieu, and B. Amanuel, *Microsatellite instability in colorectal cancer.* Asia Pac J Clin Oncol, 2010. **6**(4): p. 260-9.

58. van Engeland, M., et al., *Colorectal cancer epigenetics: complex simplicity.* J Clin Oncol, 2011. **29**(10): p. 1382-91.

59. Mohr, S., et al., *Microarrays as cancer keys: an array of possibilities.* J Clin Oncol, 2002. **20**(14): p. 3165-75.

60. Bertucci, F., et al., *Gene expression profiling of cancer by use of DNA arrays: how far from the clinic?* Lancet Oncol, 2001. **2**(11): p. 674-82.

61. Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Proc. Natl. Acad. Sci. USA, 1999. **96**: p. 6745-6750.

62. Backert, S., et al., *Differential gene expression in colon carcinoma cells and tissues detected with a cDNA array.* Int J Cancer, 1999. **82**(6): p. 868-74.

63. Hegde, P., et al., *Identification of tumor markers in models of human colorectal cancer using a 19,200-element complementary DNA microarray.* Cancer Res., 2001. **61**: p. 7792-7797.

64. Kitahara, O., et al., *Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumor tissues and normal epithelia.* Cancer Res, 2001. **61**(9): p. 3544-9.

65. Notterman, D.A., et al., *Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays.* Cancer Res, 2001. **61**(7): p. 3124-30.

66. Agrawal, D., et al., *Osteopontin Identified as Lead Marker of Colon Cancer Progression, Using Pooled Sample Expression Profiling.* J. Natl. Cancer Inst., 2002. **94**: p. 513-521.

67. Birkenkamp-Demtroder, K., et al., *Gene expression in colorectal cancer.* Cancer Res, 2002. **62**(15): p. 4352-63.

68. Lin, Y.M., et al., *Molecular diagnosis of colorectal tumors by expression profiles of 50 genes expressed differentially in adenomas and carcinomas.* Oncogene, 2002. **21**(26): p. 4120-8.

69. Zou, T.T., et al., *Application of cDNA microarrays to generate a molecular taxonomy capable of distinguishing between colon cancer and normal colon.* Oncogene, 2002. **21**: p. 4855-4862.

70. Frederiksen, C.M., et al., *Classification of Dukes' B and C colorectal cancers using expressionarrays.* J Cancer Res Clin Oncol, 2003. **129**.

71. Tureci, O., et al., *Computational dissection of tissue contamination for identification of colon cancer-specific expression profiles.* FASEB J, 2003. **17**(3): p. 376-85.

72. Williams, N.S., et al., *Identification and validation of genes involved in the pathogenesis of colorectal cancer using cDNA microarrays and RNA interference.* Clin. Cancer Res., 2003. **9**: p. 931-946.

73. Bertucci, F., et al., *Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters.* Oncogene, 0000. **23**(7): p. 1377-1391.

74. Frederiksen, C.M., et al., *Classification of Dukes' B and C colorectal cancers using expression arrays.* J Cancer Res Clin Oncol, 2003. **129**(5): p. 263-71.

75. Kwong, K.Y., et al., *Synchronous global assessment of gene and protein expression in colorectal cancer progression.* Genomics, 2005. **86**(2): p. 142-58.

76. Perez Villamil, B., et al., *Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior.* BMC Cancer, 2012. **12**: p. 260-260.

77. Cancer Genome Atlas, N., *Comprehensive molecular characterization of human colon and rectal cancer.* Nature, 2012. **487**(7407): p. 330-7.

78. Salazar, R., et al., *Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer.* J Clin Oncol, 2011. **29**(1): p. 17-24.

79. Shen, L., et al., *Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer.* Proc Natl Acad Sci U S A, 2007. **104**(47): p. 18654-9.

80. Warnat, P., R. Eils, and B. Brors, *Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes.* BMC Bioinformatics, 2005. **6**: p. 265.

81. Ghosh, D., et al., *Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer.* Funct Integr Genomics, 2003. **3**(4): p. 180-8.

82. Ramaswamy, S., et al., *A molecular signature of metastasis in primary solid tumors.* Nat Genet, 2003. **33**(1): p. 49-54.

83. Zhou, X.J., et al., *Functional annotation and network reconstruction through cross-platform integration of microarray data.* Nat Biotechnol, 2005. **23**(2): p. 238-43.

84. Stevens, J.R. and R.W. Doerge, *Combining Affymetrix microarray results.* BMC Bioinformatics, 2005. **6**: p. 57.

85. Rhodes, D.R., et al., *Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer.* Cancer Res, 2002. **62**(15): p. 4427-33.

86. Choi, J.K., et al., *Combining multiple microarray studies and modeling interstudy variation.* Bioinformatics, 2003. **19 Suppl 1**: p. i84-90.

87. Hu, P., C.M. Greenwood, and J. Beyene, *Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models.* BMC Bioinformatics, 2005. **6**: p. 128.

88. Siegel, R., D. Naishadham, and A. Jemal, *Cancer statistics, 2012.* CA Cancer J Clin, 2012. **62**(1): p. 10-29.

89. Zhang, H., et al., *Recursive partitioning for tumor classification with gene expression microarray data.* Proc Natl Acad Sci U S A, 2001. **98**(12): p. 6730-5.

90. Nannini, M., et al., *Gene expression profiling in colorectal cancer using microarray technologies: results and perspectives.* Cancer Treat Rev, 2009. **35**(3): p. 201-9.

91. Cardoso, J., et al., *Expression and genomic profiling of colorectal cancer.* Biochim Biophys Acta, 2007. **1775**(1): p. 103-37.

92. Sagynaliev, E., et al., *Web-based data warehouse on gene expression in human colorectal cancer.* Proteomics, 2005. **5**(12): p. 3066-78.

93. Chan, S.K., et al., *Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers.* Cancer Epidemiol Biomarkers Prev, 2008. **17**(3): p. 543-52.

94. Shih, W., R. Chetty, and M.S. Tsao, *Expression profiling by microarrays in colorectal cancer (Review).* Oncol Rep, 2005. **13**(3): p. 517-24.

95. Davis, S. *GEOquery R package: Get data from NCBI Gene Expression omnibus (GEO)*. 2013 [cited 2012 February]; Release (2.12):[The NCBI Gene Expression Omnibus (GEO) is a public repository of microarray data. Given the rich and varied nature of this resource, it is only natural to want to apply BioConductor tools to these data. GEOquery is the bridge between GEO and BioConductor.]. Available from: http://www.bioconductor.org/packages/2.12/bioc/html/GEOquery.html.

96. Huang, D.W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

97. Smith, J.J., et al., *Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer.* Gastroenterology, 2010. **138**(3): p. 958-68.

98. Moreno, V. and R. Sanz-Pamplona, *Altered pathways and colorectal cancer prognosis.* BMC Med, 2015. **13**: p. 76.

99. Planutis, K., M. Planutiene, and R.F. Holcombe, *A novel signaling pathway regulates colon cancer angiogenesis through Norrin.* Sci. Rep., 2014. **4**.

100. Jemal, A., et al., *Cancer statistics, 2010.* CA Cancer J Clin, 2010. **60**(5): p. 277-300.

101. Parkin, D.M., et al., *Global cancer statistics, 2002.* CA Cancer J Clin, 2005. **55**(2): p. 74-108.

102. Salhab, M., N. Patani, and K. Mokbel, *Sentinel lymph node micrometastasis in human breast cancer: an update.* Surg Oncol, 2011. **20**(4): p. e195-206.

103. Eggermont, A.M., *Adjuvant therapy of malignant melanoma and the role of sentinel node mapping.* Recent Results Cancer Res, 2000. **157**: p. 178-89.

104. Pavlidis, N., et al., *Diagnostic and therapeutic management of cancer of an unknown primary.* Eur J Cancer, 2003. **39**(14): p. 1990-2005.

105. Fernandez-Pineda, I., J.A. Sandoval, and A.M. Davidoff, *Hepatic metastatic disease in pediatric and adolescent solid tumors.* World J Hepatol, 2015. **7**(14): p. 1807-17.

106. Su, X., et al., *TAp63 suppresses metastasis through coordinate regulation of Dicer and miRNAs.* Nature, 2010. **467**(7318): p. 986-90.

107. Glud, M., et al., *Downregulation of miR-125b in metastatic cutaneous malignant melanoma.* Melanoma Res, 2010. **20**(6): p. 479-84.

108. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.

109. Calin, G.A. and C.M. Croce, *MicroRNA signatures in human cancers.* Nat Rev Cancer, 2006. **6**(11): p. 857-66.

110. Cho, W.C., *OncomiRs: the discovery and progress of microRNAs in cancers.* Mol Cancer, 2007. **6**: p. 60.

111. Rosenfeld, N., et al., *MicroRNAs accurately identify cancer tissue origin.* Nat Biotechnol, 2008. **26**(4): p. 462-9.

112. Slattery, M.L., et al., *Colorectal tumor molecular phenotype and miRNA: expression profiles and prognosis.* Mod Pathol, 2016. **29**(8): p. 915-27.

113. Schlormann, W., et al., *Influence of miRNA-106b and miRNA-135a on butyrate-regulated expression of p21 and Cyclin D2 in human colon adenoma cells.* Genes Nutr, 2015. **10**(6): p. 50.

114. Konishi, H., et al., *microRNA-26a and -584 inhibit the colorectal cancer progression through inhibition of the binding of hnRNP A1-CDK6 mRNA.* Biochem Biophys Res Commun, 2015. **467**(4): p. 847-52.

115. Wang, Y.X., et al., *Initial study of microRNA expression profiles of colonic cancer without lymph node metastasis.* J Dig Dis, 2010. **11**(1): p. 50-4.

116. Almeida, M.I., et al., *Strand-specific miR-28-5p and miR-28-3p have distinct effects in colorectal cancer cells.* Gastroenterology, 2012. **142**(4): p. 886-896 e9.

117. Rasmussen, M.H., et al., *High expression of microRNA-625-3p is associated with poor response to first-line oxaliplatin based treatment of metastatic colorectal cancer.* Mol Oncol, 2013. **7**(3): p. 637-46.

118. Matsuyama, R., et al., *MicroRNA-27b suppresses tumor progression by regulating ARFGEF1 and focal adhesion signaling.* Cancer Sci, 2016. **107**(1): p. 28-35.

119. Wang, X., et al., *Downregulation of miR-195 correlates with lymph node metastasis and poor prognosis in colorectal cancer.* Med Oncol, 2012. **29**(2): p. 919-27.

120. Tong, Z., et al., *miR-125a-5p inhibits cell proliferation and induces apoptosis in colon cancer via targeting BCL2, BCL2L12 and MCL1.* Biomed Pharmacother, 2015. **75**: p. 129-36.

121. Slattery, M.L., et al., *An evaluation and replication of miRNAs with disease stage and colorectal cancer-specific mortality.* Int J Cancer, 2015. **137**(2): p. 428-38.

122. Li, Q., et al., *miR-139-5p Inhibits the Epithelial-Mesenchymal Transition and Enhances the Chemotherapeutic Sensitivity of Colorectal Cancer Cells by Downregulating BCL2.* Sci Rep, 2016. **6**: p. 27157.

123. Okamoto, A., et al., *Enhanced Efficacy of Doxorubicin by microRNA-499-Mediated Improvement of Tumor Blood Flow.* J Clin Med, 2016. **5**(1).

124. Slattery, M.L., R.K. Wolff, and A. Lundgreen, *A pathway approach to evaluating the association between the CHIEF pathway and risk of colorectal cancer.* Carcinogenesis, 2015. **36**(1): p. 49-59.

125. Vlachos, I.S., et al., *DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways.* Nucleic Acids Res, 2012. **40**(Web Server issue): p. W498-504.

126. Qiu, C.Z., et al., *Correlation of GOLPH3 Gene with Wnt Signaling Pathway in Human Colon Cancer Cells.* J Cancer, 2016. **7**(8): p. 928-34.

127. Aminuddin, A. and P.Y. Ng, *Promising Druggable Target in Head and Neck Squamous Cell Carcinoma: Wnt Signaling.* Front Pharmacol, 2016. **7**: p. 244.

128. Oh, B.Y., et al., *Twist1-induced epithelial-mesenchymal transition according to microsatellite instability status in colon cancer cells.* Oncotarget, 2016. **7**(35): p. 57066-57076.

129. Wu, Y., et al., *Knockdown of FOXK1 alone or in combination with apoptosis-inducing 5-FU inhibits cell growth in colorectal cancer.* Oncol Rep, 2016. **36**(4): p. 2151-9.

130. Hu, Y.L., et al., *Germanicol induces selective growth inhibitory effects in human colon HCT-116 and HT29 cancer cells through induction of apoptosis, cell cycle arrest and inhibition of cell migration.* J BUON, 2016. **21**(3): p. 626-32.

131. Li, A., et al., *MicroRNA array analysis finds elevated serum miR-1290 accurately distinguishes patients with low-stage pancreatic cancer from healthy and disease controls.* Clin Cancer Res, 2013. **19**(13): p. 3600-10.

132. Wang, C., et al., *MicroRNA-323-3p inhibits cell invasion and metastasis in pancreatic ductal adenocarcinoma via direct suppression of SMAD2 and SMAD3.* Oncotarget, 2016. **7**(12): p. 14912-24.

133. Ding, H., et al., *Characterization of the microRNA expression profile of cervical squamous cell carcinoma metastases.* Asian Pac J Cancer Prev, 2014. **15**(4): p. 1675-9.

134. Huat, T.J., et al., *MicroRNA Expression Profile of Neural Progenitor-Like Cells Derived from Rat Bone Marrow Mesenchymal Stem Cells under the Influence of IGF-1, bFGF and EGF.* Int J Mol Sci, 2015. **16**(5): p. 9693-718.

135. Bi, C., et al., *Genome-wide pharmacologic unmasking identifies tumor suppressive microRNAs in multiple myeloma.* Oncotarget, 2015. **6**(28): p. 26508-18.

136. Tang, H., et al., *miR-125a inhibits the migration and invasion of liver cancer cells via suppression of the PI3K/AKT/mTOR signaling pathway.* Oncol Lett, 2015. **10**(2): p. 681-686.

137. Ninio-Many, L., et al., *MicroRNA miR-125a-3p modulates molecular pathway of motility and migration in prostate cancer cells.* Oncoscience, 2014. **1**(4): p. 250-261.

138. Choi, J.S., et al., *miRNA regulation of cytotoxic effects in mouse Sertoli cells exposed to nonylphenol.* Reprod Biol Endocrinol, 2011. **9**: p. 126.

139. Molinari, C., et al., *miR-17-92a-1 cluster host gene (MIR17HG) evaluation and response to neoadjuvant chemoradiotherapy in rectal cancer.* Onco Targets Ther, 2016. **9**: p. 2735-42.

140. Schneider, N.I. and C. Langner, *Prognostic stratification of colorectal cancer patients: current perspectives.* Cancer Manag Res, 2014. **6**: p. 291-300.

141. Derwinger, K. and B. Gustavsson, *A study of lymph node ratio in stage IV colorectal cancer.* World J Surg Oncol, 2008. **6**: p. 127.

142. Kanehisa, M., et al., *KEGG as a reference resource for gene and protein annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D457-62.

143. Fabregat, A., et al., *The Reactome pathway Knowledgebase.* Nucleic Acids Res, 2016. **44**(D1): p. D481-7.

144. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.